

# **DTLEX: Ein Expertensystem zur Objektbestimmung auf der Basis natürlichsprachiger Entscheidungsbäume**

von  
Alfred Schmitt  
Institut für Betriebs- und Dialogsysteme  
Universität Karlsruhe  
76128 Karlsruhe

## **Kurzfassung:**

Es wird über die Entwicklung eines Expertensystems zur Bestimmung von schwierig zu identifizierenden Objekten berichtet. Das System wird DTLEX = "Decision Tree and natural Language oriented Expert System" genannt. Es implementiert die besonders in der Biologie oft genutzte Technik der in natürlicher Sprache abgefaßten Bestimmungsschlüssel. Diese Bestimmungsschlüssel kann man als Entscheidungsbäume auffassen. Zusätzlich und parallel zum normalen schrittweisen Entscheidungsverfahren können aber auch Merkmale des unbekanntes Objektes direkt, also ähnlich wie beim Information Retrieval, zur Objektbestimmung genutzt werden. Diese Funktionalität erforderte eine Aufbereitung der in natürlicher Sprache beschriebenen Merkmale. Dazu wurden die Merkmalbeschreibungen automatisch mit einer maschinell verwertbaren semantischen Merkmalklassifizierung ausgestattet. Als Quelltext wurde ein umfangreicher Bestimmungsschlüssel für ca. 3100 in Europa heimische Pilze zugrundegelegt. Die gesamte Wissensbasis von DTLEX wird durch eine automatisch aus dem Quelltext erzeugte Textdatei repräsentiert, in der sich unter anderem mehr als 65000 semantisch klassifizierte Einzelmerkmale befinden.

## **Abstract:**

*DTLEX: An Expert System for the Determination of Objects on the Basis of Natural Language Decision Trees.*

We report on the development of an expert system for the determination of objects, whose identification is difficult. The system is called DTLEX = "Decision Tree and natural Language oriented Expert System". It implements the technique of keys formulated in natural language used especially in biology. These keys can be interpreted as decision trees. In addition and in parallel to the normal step by step decision process, properties and attributes of the unknown object can be used very similar to the methods used in information retrieval in order to get an object identification. This function requires a preprocessing of the properties formulated in natural language. For that purpose the properties and attributes were supplied automatically with mechanically resolvable semantic classifications. The source text used is a large key for the determination of approximately 3100 european mushrooms. The whole knowledge base of DTLEX is represented by a text file containing among other things more than 65000 semantically classified attributes.

## **1. Einleitung**

In dieser Arbeit wird über die Entwicklung eines Expertensystems zur Bestimmung von Objekten berichtet. Über das zunächst noch unbekanntes Objekt sind nur Merkmale bekannt. Man möchte dann mit Unterstützung durch das Expertensystem herausfinden, um welches Objekt es sich genau handelt. Solche Expertensysteme sind eng mit den Diagnose-Systemen z. B. für medizinische Anwendungen verwandt, bei denen die Merkmale als Krankheitssymptome und Labortestergebnisse vorliegen und

eine medizinische Diagnose gestellt werden soll. In der Systematik von Puppe [Pup88] gehören die Systeme zur Objekterkennung in die Klasse der Diagnosesysteme.

Ausgangspunkt dieser Arbeit sind sogenannte dichotome Bestimmungsschlüssel, die insbesondere in der Biologie verbreitet sind. Solche Bestimmungsschlüssel dienen dazu, Exemplare einer zunächst noch unbekanntem Art oder Gattung anhand der festgestellten Merkmale und Untersuchungsbefunde zu identifizieren. In unserem Fall verwenden wir den in der Mykologie (= Pilzkunde) wohlbekanntem Bestimmungsschlüssel vom Meinhard Moser [Mos83], den wir im folgenden einfach den "Moser" nennen. (Dem Gustav Fischer Verlag bin ich zu herzlichem Dank verpflichtet, weil er mir gestattet hat, den Buchtext für diese Forschungsarbeit zu verwenden.) Er enthält praktisch alle in Europa vorkommenden Röhrlinge und Blätterpilze, also 173 Gattungen mit zusammen ca. 3100 Arten. Im Vergleich zu vielen sonst bekanntgewordenen rechnergestützten Bestimmungssystemen für Pilze ist dies ein großer Artenumfang, unser System ist also mit einer umfassenden Wissensbasis ausgestattet. Dieser Datenumfang erfordert für den Wissenserwerb eine weitgehende Automatisierung, weil eine Aufbereitung der Daten von Hand viel zu aufwendig wäre.

Daraus kann bereits das erste Ziel dieser Arbeit abgeleitet werden. Es sollte untersucht werden, wie sich das in vielen veröffentlichten Bestimmungsschlüsseln dokumentierte Expertenwissen halb- oder vollautomatisch erschließen läßt. Da dieses Wissen in natürlicher Sprache dokumentiert ist, mußte auch das Problem der automatisierten Merkmalextraktion aus natürlichsprachigen Texten angegangen und gelöst werden. Wenn man nun das in einem Bestimmungsschlüssel steckende Wissen aufbereitet und verarbeitbar gemacht hat, so ist noch keineswegs klar, wie das Expertensystem mit dem Benutzer kommuniziert bzw. wie der Dialog zur Bestimmung von Objekten abläuft. Dies sollte dann in einem zweiten Schritt des Projektes untersucht werden.

Diese Arbeit hat Querbeziehungen zu Dokumentationssystemen (Information Retrieval), die heute vor allem als Bibliotheks-Informationssysteme eingesetzt werden. In gewöhnlichen Dokumentationssystemen wird mit Hilfe von Anfragen nach passenden Objekten gesucht, in der Regel mit gezielt gewählten Stichwort-Kombinationen, vgl. z. B. [SG83]. Ähnliche Funktionalität haben auch Hypercard- bzw. Hypertext-Systeme. Was diesen Systemen normalerweise fehlt, ist eine hierarchische Struktur des Datenbestandes und die Navigationsführung durch problemangepaßte Entscheidungsbäume, mit der wir uns hier wesentlich auseinandersetzen werden.

Die Ergebnisse und Methodik der hier entwickelten Systemstruktur sind verallgemeinerungsfähig. Mit der vorgestellten Methode kann jeder Bestimmungsschlüssel in ein Expertensystem umgesetzt werden. Denn der einzelne Schlüssel stellt zunächst nur eine allerdings bereits gut strukturierte Wissensbasis dar, die letztendlich den bei realen Anwendungen stets problematischen Wissenserwerb erheblich vereinfacht. In der Biologie sind die Bestimmungsschlüssel das Ergebnis langer und ausgedehnter Forschungstätigkeiten, in die fast das gesamte veröffentlichte Wissen des engeren Fachgebietes einfließt. Sie sind gewissermaßen ein auf die Objektidentifizierung bzw. Artabgrenzung konzentriertes Wissenskondensat. Es wird später noch genauer begründet, daß durch die Baumstruktur eine erhebliche Verdichtung der insgesamt im Schlüssel steckenden Informationsmenge erzielt wird.

Nach der Problemanalyse-Phase für unser Projekt schälten sich folgende Arbeitsschwerpunkte heraus, über die hier näher berichtet wird:

- (1) Aufbereitung eines größeren natürlichsprachig vorliegenden Quelltextes (ca. 1,4 MByte) für die automatisch vorzunehmende Umsetzung in ein geeignetes Zwischenformat mit strenger syntaktischer Gliederung.
- (2) Automatische Erkennung und Klassifizierung von Einzelmerkmalen aus natürlichsprachigen Merkmalbeschreibungen.
- (3) Übersetzung der mit (1) und (2) gewonnenen Daten in ein allgemeines, anwendungsunabhängiges Datenformat, das "Stammdatei-Format" genannt wird.
- (4) Architektur und Gestaltung des eigentlichen Expertensystems, also der Shell, das der Anwender im Dialog benutzt und dessen Wissensbasis aus einer Stammdatei geladen wird.

Die Punkte (1) bis (3) dienen sämtlich dem Wissenserwerb, während (4) die eigentliche innere Logik des Expertensystems in anwendungsorientierter Verpackung beinhaltet.

Die Entwicklung des DTLEX-Systems (DTLEX = Decision Tree and natural Language oriented Expert System) hat der Autor von Januar bis September 1995 im Stile der schnellen Prototyp-Entwicklung ("rapid prototyping") durchgeführt. Zwar wurden Hilfsmittel wie z. B. PROLOG und LISP auf ihren Nutzen hin geprüft, am Ende wurde jedoch entschieden, alles maßgeschneidert in Pascal zu programmieren und auch die Datenbestände in selbststrukturierten Dateien zu halten. An Werkzeugen wurden demgemäß auf einem Apple Powerbook 520 die Programmumgebung Think-Pascal der Symantec Corp. sowie der Text-Editor BEdit eingesetzt. Insbesondere der genannte Editor, eine Entwicklung der Bare Bones Software Inc., hat sich dabei sehr bewährt, weil er vergleichsweise schnell ist und auch Dateien von 5 Mbyte Umfang bewältigt, was in vielen Fällen zwecks Arbeitersparnis die beste Lösung war.

Das Bestimmen von Pilzen ist selbst für Spezialisten manchmal sehr problematisch. So wurde z. B. von einem Treffen der Cortinarien-Experten [Kou93] über die während der Tagung bei Exkursionen aufgesammelten Pilze berichtet: "Häufig gefundene Arten: Cortinarius mucifloides, C. obtusus, C. torvus. Seltener: C. anthracinus, C. bavaricus, C. ochraceus. Einige konnten nicht bestimmt werden. Über viele Hydrocyben und Telamonien gab es lange Diskussionen." Man kann daraus aber auch folgern, daß man von Objektbestimmungssystemen keine Wunderdinge erwarten sollte, denn was nicht in der Wissensbasis steckt, kann das System auch nicht liefern. Außerdem kann z. B. DTLEX nur von Benutzern mit einer gewissen mykologischen Vorbildung und Erfahrung in der Merkmalbeurteilung sinnvoll eingesetzt werden. (Es eignet sich definitiv nicht für normale Pilzsucher, die den Wald nach eßbaren Arten durchkämmen.) Ein System wie DTLEX kann, von ernsthafteren Mykologen eingesetzt, manche Bestimmungsvorgänge aber erheblich erleichtern und beschleunigen. Insbesondere scheint es auch für die akademische Grundausbildung von Biologen geeignet zu sein, bei denen Bestimmungsübungen auf dem Studienplan stehen.

## **2. Dichotome und synoptische Bestimmungsschlüssel**

In der Biologie ist es seit langem üblich, Arten, Gattungen etc. bezüglich ihrer Merkmale in sogenannte dichotome Bestimmungsschlüssel einzuordnen. Solche Bestimmungsschlüssel sind nach festen formalen Regeln aufgebaut.

Da nicht unterstellt werden kann, daß Nichtbiologen mit der Technik und Methodik der Bestimmungsschlüssel vertraut sind, soll diese hier genauer beschrieben werden. Allgemein wird nach folgenden Prinzipien vorgegangen: Ein großer Bestimmungsschlüssel besteht aus einer größeren Zahl von Teilschlüsseln, die oft nur wenige Zeilen bis zu einigen Seiten Umfang haben. Diese Teilschlüssel sind aber untereinander hierarchisch vernetzt. Der Bestimmungsvorgang startet beim obersten Schlüssel der Hierarchie, dem Hauptschlüssel. Von diesem wird man in der Regel auf Teilschlüssel verwiesen. Der Abstieg in feiner auflösende Teilschlüssel kann sich mehrmals wiederholen, bis man schließlich auf das zu bestimmende Objekt stößt und es identifizieren kann. Der Bestimmungsvorgang verläuft also grundsätzlich vom Allgemeinen zum Speziellen. In der Biologie lehnt sich die hierarchische Gliederung gewöhnlich an die für das jeweilige Gebiet der Organismen erarbeitete Systematik an, wobei man z. B. zwischen Abteilungen, Klassen, Ordnungen, Familien, Gattungen, Untergattungen/Sektionen, Untersektionen und schließlich Arten (Species) als Endglieder unterscheidet. Sektionen und Untersektionen sind nur bei artenreichen Gattungen eingerichtet.

Zwei Beispiele von konkreten Schlüsseltexten sollen die Technik der dichotomen Bestimmungsschlüssel näher erläutern. In Textbeispiel 2.1 ist der Hauptschlüssel des in deutschsprachigen Ländern sehr bekannten Bestimmungswerkes von Schmeil-Fitch [SF76] wiedergegeben. Dieses Bestimmungswerk erschien erstmals um 1900 und erlebte bis heute 89 Auflagen mit etlichen, dem jeweils neuesten Kenntnisstand folgenden Überarbeitungen.

**Textbeispiel 2.1:** Hauptschlüssel von Schmeil-Fichten: "Flora von Deutschland und seinen angrenzenden Gebieten" [SF76]

---

**A. Tabelle zum Bestimmen der Hauptgruppen**

1.	Pfl. mit Bltn., die entw. Sttblätt. od. Frblätt. od. beide enthalten u. Samen hervorbringen.....	3
—	Pfl. stets ohne od. nur selten mit Bltn. u. Samen .....	2
2.	Pfl. nicht in Stg. u. Blätt. geglied., aus linsenf., bis 1,5 mm großen, wurzellosen ( <i>Wolffia</i> , 147) od. rundl.-längl., bis 15 mm großen, auf der Wasseroberfläche schwimmenden od. submersen, wurzelnden Gliedern bestehend (144-146) .....	<b>Lemnaceae</b> , Seite 491
—	Pfl. stets in Stg., Blatt u. Wurzel geglied., vorwgd. Land-, seltener Sumpfu. Wasserpfl.; Vermehrung durch mikroskopisch kleine, einzellige Sporen, die in besonderen Behältern, den Sporangien (2d, Sp), entstehen <i>Sporen-, Farnpfl., Pteridophyta, Tabelle I</i> , Seite 37	
3(I).	Samen frei, nicht in einen Frkn. eingeschlossen (85d); Bäume u. Sträucher, mit nadelf. od. schuppenf., meist immergrünen Blätt. (338-340) <i>Nacktsamige Pfl., Gymnospermae, Tabelle II</i> , Seite 38	
—	Samen in einen Frkn. eingeschlossen (85c, 86-89) <i>Bedecktsamige Pfl., Angiospermae</i> .....	4
4.	Pfl. (Kräuter u. Stauden) z. Bltzt. ohne grüne Blätt. od. Bltnstg. mit bleichen, bräunl. od. violetten Schuppenblätt. ....	<b>Tabelle III</b> , Seite 39
—	Pfl. z. Bltzt. mit wohl entwickelten grünen Blätt. ....	5
5.	Blattspreite streifen-(parallel-) od. bogennervig, selten netz- od. l-nervig, stets einfach u. ungeteilt, zuw. stielrund ( <i>Juncus</i> ) od. nadelf. ( <i>Asparagus</i> ) ausgebildet, häufig in 2 Zeilen od. in 2- bis mehrzähligen Quirlen; Sproß stets krautig, auf dem Querschnitt mit zerstreut angeordneten Leitbündeln (Lupe! 125a); Knollen, Rhizome, Zwiebeln ohne hervortretende Hauptwurzel als ausdauernde Organe vorherrschend; Bltnorgane meist in 3-zähligen Wirteln; Bltnhülle selten in Kelch u. Blkr. gegliedert (125b); Keimling nur mit 1 Keimblatt ( <i>lc</i> ). <i>Einkeimblättrige Pfl., Monocotyledoneae, Tabelle IV</i> , Seite 39	
—	Blattspreite mit fiedrig, fingerig, handf. od. netzart. miteinander verbundenen Nerven, seltener bogen-, parallel- od. l-nervig (zuw. erst nach der Blüte erscheinend), einfach, gefied., gefing., zuw. mit Ranken; Blätt. zerstreut, gekreuzt-gegenst. od. in mehrzähligen Wirteln, seltener in 2 Zeilen; Sproß krautig od. holzig, auf dem Querschnitt mit ringf. angeordneten Leitbündeln (Lupe! 126a); Bltnorgane meist in 5-zähligen Kreisen; Bltnhülle häufig in K. u. Blkr. geschieden (126b); Keimling mit 2 Keimblättern ( <i>lb</i> ) <i>Zweikeimblättrige Pfl., Dicotyledoneae, Tabelle V</i> , Seite 42	

---

Der Schlüssel hat neben der Überschrift keinen einleitenden Kopftext und besteht aus 5 sogenannten Nummern. Es handelt sich um den Hauptschlüssel, das heißt, jede Bestimmung einer Pflanze startet bei diesem Hauptschlüssel. Es fällt auf, daß bei der sprachlichen Schilderung von Merkmalen in ganz erheblichem Umfang Abkürzungen verwendet werden, um den Text zu kürzen, die Seitenzahl des Gesamtwerkes zu reduzieren und damit auch die Tauglichkeit für Exkursionen zu erhöhen, bei denen das Bestimmungswerk (z. B. in Kleidungstaschen) mitgeführt werden soll. Die Bestimmung startet bei Nummer 1. Nummern sind grundsätzlich zweigeteilt, das erklärt den Begriff dichotom = zweigeteilt. Der zweite Teil der Nummer beginnt jeweils nach dem herausgerückt gedruckten "—". Ist man beim Bestimmen an einer Nummer angelangt, so muß man die Merkmalbeschreibungen beider Teile prüfen und sich dann letztendlich entscheiden, welche Merkmale auf das zu bestimmende Objekt zutreffen. Vor allem bei kritischeren Entscheidungen ist es von Vorteil, zwischen beiden Möglichkeiten sorgfältig abwägen zu können. Sodann folgt man dem Verweis, der am Ende des zutreffenden Textteiles folgt. Dieser ist entweder eine andere Nummer (wie oben z. B. bei Nummer 1

die Verweise auf die Nummern 3 und 2) oder ein Verweis auf einen Unterschlüssel. Die global angesiedelten Schlüssel werden im Schmeil-Fitchen "Tabellen" genannt, während z. B. der erste Verweis bei Nummer 2 auf die Familie Lemnaceae verweist. Der dort auf Seite 491 zu findende Schlüssel verweist seinerseits auf die Gattungen innerhalb der Familie und bei den Gattungsschlüsseln findet man schließlich Arten. Durch die zusätzliche Angabe von Seitenzahlen wird das Verweisziel auch in einem dickeren Buch schnell gefunden.

Das zweite Textbeispiel bezieht sich auf den Moser [Mos83], hier wird der Teilschlüssel für die Gattung *Boletus* zusammen mit dem Anfang des anschließenden Unterschlüssels für die Sektion *Luridi* gezeigt.

## Textbeispiel 2.2:

---

#6538:2.2.10 *Boletus* Dill. ex Fr., Röhrlinge@  
#6539:{Röhrlinge im engeren Sinne}.  
#6540:Frkp. fleischig, St. meist dickbauchig (Ausnahme 2.2.10.2.1), mit feinflockiger oder  
#6541:netziger Oberfläche. H. sehr bald trocken, fein filzig oder kahl. Spp. oliv bis olivbraun.  
#6542:Sp. spindelig oder fast spindelig, glatt. Rö.-Tr. mit stark divergierenden, locker verflochtenen,  
#6543:± hyalinen oder blassen Hyphen in den Seitenschichten, Rö.-Tr. im Mediostratum  
#6544:satter gefärbt (wenn pigmentiert). Pigmente bisher Variegatsäure, Variegatorubin  
#6545:und Xerocomsäure bekannt.  
#6546:1. Hymenophor mit ± roten, rotorange, orange, orangegelben (und  
#6547:später etwas rot werdenden) Po. 2.2.10.1  
#6548:1\*. Hymenophor mit weißen oder ± gelben Po. 2  
#6601:2. Hymenophor mit von Jugend an gelben Po. 2.2.10.2 (S. 67)  
#6602:2\*. Hymenophor jung weißlich, später erst gelb, grünlichgelb verfärbend  
#6603:2.2.10.3 (S. 68)  
#6604:\$  
#6605:2.2.10.1 Sekt. *Luridi* Fr.@  
#6606:Poren orange bis rot.  
#6607:1. St. ohne Netz, kahl oder feinflockig oder schuppig 2  
#6608:1\*. St. mit deutlicher, netzartiger Zeichnung 5  
#6609:2. H. schleimig, dann trocken. Der ganze Frkp. ± satt karmin-blutrot. H. u. St.  
#6610:mit gelber Grundfarbe. H. 7-10 cm, St. feinflockig. Fl. blaß gelb, blauend. Sp.  
#6611:10-18/4-6 mü. Laubwald u. Mischwald. - KM 406, BSMF 47, Atl. 44, Sing.  
#6612:VI 13, C 268, PD 60 a-d, Md 157, LE 51 2.2.10.1.1 *B. dupainii* Boud.  
#6613:2\*. H. von Anfang an trocken. Po. mit mehr braunroten oder orangeroten  
#6614:Farben (Hexenröhrlinge) 3  
#6615:3. H. grauweißlich, grauockerlich, dann rot oder purpurn gefleckt,  
#6616:an dem Rd. auch satt gelb, filzig. St. auf gelbem Grunde blaß rötlich, später  
#6617:bräunlich flockig-punktiert. Fl. gelb, stark blauend. Sp. 12-17,5/5-6  
#6618:müm. Unter Buchen. - BSMF 64, Atl. 90, Sing. VI 16, PD 65 a-c, C  
#6619:704, Md 159, LE 54 2.2.10.1.2 *B. lupinus* Krbh., Wolfsröhrling  
#6620:3\*. H. mit dunklen zumindest teilweise ± braunen Farben 4  
#6621:4. H. ± dunkelbraun, samtig, 5-20 cm. {gelbhütige Form bekannt:  
#6622:ssp. *discolor* (Quel.) Derm., Kuth. et Sing. }. St. auf gelblichem  
#6623:Grunde fein karminrot flockig. Po. rot (braun). Fl. zitronengelb,  
#6624:Schnitt sofort dunkelblau. Sp. 11-19/4,5-7 mü. Laubwald u. Nadelwald.  
#6625:Bes. kalkarme Böden. - Kbch. 30, MH I 157, MP 69, SPT II 60, Sing.  
#6626:VI 14, R 130, C 276, PD 60 e-h, 61, Md 61, LE 47  
#6627:2.2.10.1.3 \* *B. erythropus* (Fr. ex Fr.) Pers.  
#6628:(= *miniatoporus* Secr.), Schusterpilz, Donnerpilz.  
#6629:4\*. H. braun bis oliv oder orangerot, seltener dunkelrot. St. gelblich bis  
#6630:braunrot, Basis dunkelrot (auch im Fl.). Fl. schwach blauend. Po.  
#6631:mehr orangerot. Sp. 10-17/5-8 mü. Laubwald. - Kbch. 8, BSMF 51,  
#6632:Atl. 66, 67, CM 7, Tf. 10, R 128, SPT IV 51, Sing. VI 14, 15, PD  
#6633:59, C 701, Md 162, LE 49, FRIC 1 2.2.10.1.4 \* *B. queletii* Schulz.  
#6634:(= *erythropus* Pers. = *bresadolae* Schulz.),  
#6635:Glattstieliger Hexenröhrling

#6636:5. Netz grob u. langgezogen. St. meist ± orange, Basis weinrot. H.  
#6637: von olivbraun, orangebraun bis dunkel-braun, auch rötlich. Fl. blaß, in Basis oft  
#6638: weinrötlich, schwach blauend. Po. rotbraun. Sp. 9-17/5-7 müm. Laubwald  
#6639: u. Nadelwald. - Kbch. 15, 16, MH I 158, Ha 4, MP 70, SPT II 59, R  
#6640: 129, Sing. VI 19, 18, C 275, 702, PD 62, 63, Md 160, LE 46  
#6641: 2.2.10.1.5 (\*?) B. luridus Schff. ex Fr., Netzstieliger Hexenröhrling  
#6642: 5\*. Netzmaschen fein, nicht langgezogen 6  
usw. usf.

---

Abgedruckt ist der Text von Seite 65, Zeile 38 (#6538) bis Seite 66, Zeile 42 (#6642), und zwar nicht mehr im Druckbild der Buchform, sondern in der Form der für die Weiterverarbeitung leicht aufbereiteten Computer-Textdatei. Der Text beginnt mit der Gattung "Boletus". Es folgt ein Kommentar in geschweiften Klammern {Röhrlinge im engeren Sinne}, sodann die Beschreibung der Merkmale, durch die die Gattung charakterisiert ist. Z. B. sind die Sporen (Sp.) bei Arten der Gattung Boletus spindelrig oder fast spindelrig und glatt. (Da die Sporen je nach Pilzart Ausdehnungen zwischen 3 und 20 µm haben, kann ein solches Merkmal nur durch hochauflösende mikroskopische Untersuchung geprüft werden.) Nach dieser Beschreibung folgt bis zum Ende-Zeichen \$ in Zeile #6604 ein Nummernteil, in dem Verweise auf die drei Sektionen 2.2.10.1, 2.2.10.2 und 2.2.10.3 erfolgen, in die die Gattung Boletus = 2.2.10 zerlegt ist. Im Moser wird der zweite Teil einer Nummer mit Nummer und Stern "\*" gekennzeichnet.

Wenn der Bestimmer auf eine Nummer verwiesen wird, so prüft er genau wie bei jedem dichotomen Bestimmungsschlüssel die Merkmale des Nichtstern-Teils und des Stern-Teils auf Übereinstimmung mit dem zu bestimmenden Objekt und folgt dem Verweis, der zu den zutreffenden Merkmalen gehört. Ab Zeile #6605 folgt der Anfang des Unterschlüssels für die Sektion 2.2.10.1, die "Luridi" genannt wird und bereits von dem berühmten schwedischen Mykologen Fr. = Fries (1794-1878) vorgeschlagen wurde. Der Schlüsseltext der Sektion ist nur bis zur Nummer 5 abgedruckt, geht im Moser aber bis zur Nummer 9. In diesem Schlüssel finden sich jetzt auch sogenannte Artausschlüsselungen, z. B. in den Zeilen #6612, #6619 und #6627. Artausschlüsselungen sind Endknoten, wenn man an einem solchen ankommt und die Merkmale stimmen mit dem unbekanntem Objekt überein, so war die Bestimmung erfolgreich, die ausgeschlüsselte Art ist das Bestimmungsergebnis. Treten bei der Bestimmung Widersprüche auf, z. B. dadurch, daß weder der Nichtstern-Teil noch der Stern-Teil zutreffende Merkmale vorgibt, so muß die Bestimmung ergebnislos abgebrochen werden. Normalerweise geschieht dies bei nicht korrekten Entscheidungen in vorangegangenen Schritten, bei falscher Beurteilung von Merkmalen usw. In seltenen Fällen kann es auch vorkommen, daß die zu dem unbekanntem Objekt gehörende Art im Schlüssel nicht repräsentiert ist oder überhaupt der Wissenschaft noch unbekannt ist. (Der Moser enthält nicht alle Großpilze Europas, sondern nur die Arten der vier Ordnungen 1. Polyporales, 2. Boletales, 3. Agaricales und 4. Russulales. Für die restlichen ca. 4000 Großpilze gibt es weitere Bestimmungsschlüssel, die mit dem hier entwickelten Verfahren ebenfalls in ein Bestimmungssystem umgesetzt werden könnten, insbesondere auch in ein gemeinsames Bestimmungssystem für alle Großpilze.)

Der Moser-Schlüssel handhabt die Verweise zwischen Teilschlüsseln etwas anders als das sonst üblich ist, denn es wird eine Dezimalklassifikation verwendet, um Ordnungen, Familien, Gattungen, Sektionen und ganz zum Schluß auch Arten eindeutig zu kennzeichnen. Der in Zeile #6627 ausgeschlüsselte "Boletus erythropus" hat die Klassifikationsnummer 2.2.10.1.3, er ist eingeordnet in die Sektion 2.2.10.1 Luridi, diese gehört zur Gattung 2.2.10 Boletus, welche ihrerseits in der Familie 2.2 Boletaceae untergebracht ist, die wiederum zur Ordnung 2. Boletales gehört. Man vermutet sofort, daß dies für die Implementierung des Expertensystems und besonders die Strukturierung der Daten eine bedeutende Hilfe ist. Das hat sich auch bestätigt. Vorteile ergaben sich dadurch, daß diese Klassifikationsnummern bei der syntaktischen Analyse des Gesamttextes als elementare Zerteilungs-Objekte unverwechselbar erkannt werden können. Das vereinfacht die Adressierung innerhalb des Gesamtschlüssels erheblich. Schließlich wird dieses den Benutzern des Moser wohlvertraute Nummerierungssystem in der Dialogschnittstelle benutzt, um Suchraumeinschränkungen z. B. auf eine bestimmte Gattung rasch vornehmen zu können.

Die Anwendung dichotomer Bestimmungsschlüssel in der Praxis funktioniert nur selten ohne Probleme:

- (1) Die Merkmale sind häufig unscharf definiert, so daß Fehleinschätzungen nicht selten sind. Das gilt für Farben, Formen und auch für Norm-Maße, die im Einzelfall deutlich über- oder unterschritten werden können.
- (2) In manchen Fällen sind 40 und mehr Einzelschritte (Nummern) zu durchlaufen, um die zutreffende Art zu erreichen. Das erfordert eine sehr hohe Konzentration. Eine einzige Fehlentscheidung führt bereits in die Irre und kann später nicht mehr korrigiert werden.
- (3) Manche aufgesammelten Pilze zeigen einzelne arttypische Merkmale nur undeutlich.

Am wenigsten befriedigt an den dichotomen Schlüsseln, daß der Bestimmer die Merkmale in vorgegebener Reihenfolge und dann auch noch lückenlos parat haben muß.

### **Synoptische Bestimmungsschlüssel, bisher bekanntgewordene Pilzbestimmungssysteme**

Viel attraktiver scheint folgende Vorgehensweise: Man bietet dem Expertensystem die festgestellten Merkmale des unbekanntes Objektes an und fragt nach allen Arten, die ebenfalls diese Merkmale haben. Denn wenn man schon eine Anzahl Merkmale hat, kann es sein, daß durch sie oder bereits durch eine Teilmenge davon das zu bestimmende Objekt eindeutig feststeht. Diese Vorgehensweise liegt den sogenannten synoptischen Bestimmungsschlüsseln zugrunde. Sie sind allerdings nicht mehr in Buchform, sondern nur noch im Computer realisierbar. Heute nicht mehr aktuell sind auch frühere Versuche mit individuell gelochten Merkmalkarten. Dobbitsch [Dob88] weist auf die Schwierigkeiten beim praktischen Arbeiten mit dichotomen Schlüsseln hin und plädiert für die synoptische Vorgehensweise. Er hat auch ein erstes synoptisch vorgehendes Bestimmungsprogramm geschrieben. Mendt u. Georgi [MG90] bemängeln, daß bei strikter Handhabung der synoptischen Methode leicht folgendes geschieht: Man hat z. B. 9 von 10 Merkmalen richtig vorgegeben, eines jedoch falsch. Die zutreffende Art wird dann ausgeschlossen und nicht erkannt. Sie schlagen daher ein flexibleres Bewertungssystem vor, indem sie mit einem Bonuspunktesystem zählen, wie viele Merkmalübereinstimmungen die einzelne Art erzielt. Arten mit hoher Punktzahl weisen dann auch eine hohe Übereinstimmung mit den vorgegebenen Merkmalen auf. Dieses Verfahren ist noch erweiterbar, indem der Benutzer zwischen wichtigen und unwichtigen Merkmalen unterscheidet. Ein neueres Pilzbestimmungssystem SYNOPSE von Schilling [Sch95] nutzt die Gewichtungstrategie und erzielt mit relativ detaillierten Merkmalmengen sehr ansprechende Bestimmungsleistungen. Einen etwas anderen Ansatz verfolgt das Bestimmungssystem PIDIAS von Neis [Nei88]. Er setzt die Expertensystem-Shell MED2 von Puppe [Pup86] ein, bei dem ein System von Folgerungsregeln der Gestalt "Merkmal -> Diagnose" dabei hilft, Pilze zu bestimmen. Leider verfügen alle genannten Systeme nur über relativ kleine Datenbasen für wenige hundert Arten, weil die Erstellung der Merkmalzuordnungen in Handarbeit erfolgen mußte und daher sehr arbeitsaufwendig ist.

### **Die DTLEX-Strategie**

Bei ersten Versuchen mit Vorversionen von DTLEX zeigte sich, daß auch der synoptische Ansatz insbesondere bei größeren Objektmengen seine Probleme hat, die durch den unscharfen Charakter der meisten Merkmale verursacht werden. Schwerwiegender ist jedoch der Umstand, daß für ein vorgegebenes Merkmal praktisch für jeden Pilz korrekt entschieden werden muß, ob es zutrifft oder nicht. Für den einzelnen Pilz sind aber nur relativ wenige Merkmale charakteristisch, während die gesamte Merkmalmenge über alle Pilze genommen um ein Vielfaches größer ist. In den dichotomen Bestimmungsschlüsseln werden vorzugsweise nur solche Merkmale genutzt, die unterscheidenden Charakter haben. Daher ist die Merkmalmenge nicht vollständig und ist in dieser Form für die synoptische Vorgehensweise nur bedingt brauchbar. Für DTLEX wurden aus ersten ernüchternden Erfahrungen mit der synoptischen Vorgehensweise folgende Konsequenzen gezogen:

(1) Der normale Bestimmungsvorgang mit dem dichotomen Schlüssel ist die Basis des Bestimmungsverfahrens, weil nur so korrekte, von der Fachwelt anerkannte Ergebnisse erzielt werden, insbesondere wenn es um subtilere Unterscheidungen geht. Bei sinnvoller Anwendung sind die Bestimmungsleistungen des Systems also mindestens so gut wie bei Verwendung des Buch-Schlüssels.

(2) Durch zusätzliche, vom Anwender nach eigener Wahl vorgebbare Merkmaleingaben soll parallel zum Bestimmungsschlüssel-Verfahren eine merkmalorientierte Objektidentifikation unterstützt werden, mit der auch im synoptischen Direktverfahren Pilze bestimmt werden können. So kann man sich schnell einen Überblick verschaffen, welche Pilze grobe Merkmalübereinstimmung zeigen.

Angestrebt wird also eine Symbiose zwischen der Bestimmungsschlüssel-Methode und der merkmalorientierten Objektsuche.

### 3. Formale Charakterisierung dichotomer Bestimmungsschlüssel

Der Aufbau von dichotomen Bestimmungsschlüsseln ist aus den oben besprochenen Beispielen ersichtlich geworden. Vom allgemeinen Schema her ist ein einzelner Schlüssel also stets wie folgt gegliedert:

Kopftext: Überschrift, Merkmale

1.	Merkmale	Verweis
1*.	Merkmale	Verweis
2.	Merkmale/Artausschlüsselung	(kein Verweis)
2*.	Merkmale	Verweis
3.	Merkmale	Verweis
3*.	Merkmale	Verweis
4.	Merkmale/Artausschlüsselung	(kein Verweis)
4*.	Merkmale	Verweis
5.	Merkmale	Verweis
5*.	Merkmale	Verweis
6.	Merkmale/Artausschlüsselung	(kein Verweis)
6*.	Merkmale/Artausschlüsselung	(kein Verweis)
7.	Merkmale/Artausschlüsselung	(kein Verweis)
7*.	Merkmale	Verweis
8.	Merkmale	Verweis
8*.	Merkmale	Verweis
9.	Merkmale	Verweis
9*.	Merkmale	Verweis
10.	Merkmale/Artausschlüsselung	(kein Verweis)
10*.	Merkmale/Artausschlüsselung	(kein Verweis)

Jeder Teilschlüssel beginnt mit einem Kopftext, der aus der Überschrift des Teilschlüssels und den gemeinsamen Merkmalen für alle Objekte besteht, die in dem Teilschlüssel weiter aufgespalten werden. Die Nummern sind grundsätzlich gesplittet in einen ersten und zweiten Teil, den Nichtstern- und Stern-Teil. Verweise erfolgen entweder auf Nummern (stets ohne Stern) oder auf andere Teil-Bestimmungsschlüssel, die dem aktuellen Schlüssel dann in der Regel untergeordnet sind.

#### Definition des Adressierungsschemas

Um Bestimmungsschlüssel in eine für die Implementierung geeignete Struktur zu bringen, muß zunächst ein Adressierungsschema entwickelt werden. Wir definieren daher:

**Adressen** sind Tripel (Schlüsselkennzeichen, Vorzeichen, Nummerindex).

**Schlüsselkennzeichen** identifizieren den jeweiligen Teilschlüssel des Gesamt-Bestimmungsschlüssels.



Das *Vorzeichen* wird aus der Menge  $\{+,-\}$  genommen.  
Der *Nummernindex* ist eine Zahl aus  $\{0, 1, 2, 3, \dots\}$ .

Insbesondere werden alle Verweise innerhalb der Schlüssel durch Adressen repräsentiert. Im Falle des Moser verwenden wir die in Abschnitt 2 erläuterte Dezimalklassifikation, um die Teilschlüssel eindeutig mit Schlüsselkennzeichen zu belegen. Die Adresse für den Teilschlüssel "Boletus" lautet also (2.2.10,+0), oder einfach als Zeichenkette "2.2.10+0" geschrieben. Der Kopftext eines Teilschlüssels hat also den Nummernindex 0 mit Vorzeichen +. Analog dazu hat der Teilschlüssel für die Sektion Luridi die Adresse 2.2.10.1+0. Die Nummernindices 1, 2, ... werden für die Unteradressierung der einzelnen Nummern verwendet, wobei das Vorzeichen "-" den Stern-Teil der Nummer, das Vorzeichen "+" den Nichtstern-Teil spezifiziert. Die Nummer 5 im Teilschlüssel Luridi hat also die Adresse 2.2.10.1+5, während der zugehörige Stern-Teil 5\* die Adresse 2.2.10.1-5 hat. Im Moser haben die oberhalb der Gattungen angesiedelten Schlüssel die Schlüsselkennzeichen A, B, ..., P. Dem Hauptschlüssel wurde das Schlüsselkennzeichen @ zugeordnet. Daher ist @+0 die sogenannte Startadresse des Gesamtschlüssels, bei der jede Bestimmung beginnt.

### Definition des Entscheidungsbaumes

Im nächsten Schritt können wir nun die Struktur des Entscheidungsbaumes definieren bzw. formalisieren, die sich aus dem gesamten Bestimmungsschlüssel ergibt:  
Jeder Kopftext, jeder Nichtstern-Nummernteil und jeder Stern-Nummernteil des gesamten Bestimmungsschlüssels wird im Entscheidungsbaum durch einen *Knoten* repräsentiert. Die Menge der Knoten liegt dadurch eindeutig fest. Jeder Knoten hat, wie oben eingeführt, eine eindeutige *Adresse*. Der Kopftext erhält einen Positiv-Verweis auf die (Nichtstern-)Nummer "1.". Knoten, die keine Artausschlüsselung repräsentieren, haben einen *Positiv-Verweis* und evtl. einen *Negativ-Verweis*. Die Positiv-Verweise sind genau die im gedruckten Schlüssel ausgewiesenen Verweise. Die Negativ-Verweise werden zusätzlich eingeführt und zeigen vom Nichtstern-Nummernteil auf den zugehörigen Stern-Nummernteil. Dadurch wird das bei einer einzelnen Nummer anzuwendende Entscheidungsverfahren formal in zwei Einzelschritte aufgelöst.  
Der Entscheidungsbaum hat einen ausgezeichneten *Startknoten*, bei dem jede Bestimmung beginnt. Dies ist der Kopftext des Hauptschlüssels, der wie bereits erwähnt die Adresse @+0 hat. Alle anderen Teilschlüssel sind dem Hauptschlüssel im Sinne der Hierarchie untergeordnet.

Damit ist nun hinreichend geklärt, wie aus jedem dichotomen Bestimmungsschlüssel formal ein Entscheidungsbaum abgeleitet werden kann. Im Falle des Moser hat der Baum über 8000 Knoten. Streng genommen ist es kein Baum, sondern nur ein gerichteter Graph ohne Zyklen, denn es gibt vereinzelt Knoten, auf die von mehreren Stellen verwiesen wird. Wir benutzen aber weiterhin den Begriff **Entscheidungsbaum**, weil er das zugrundeliegende Konzept sehr gut verdeutlicht.

Eine herausragende Eigenschaft der Entscheidungsbäume ist die **kompakte Darstellung** von Merkmalinformationen. Wir betrachten z. B. eine Artausschlüsselung und den Weg vom Startknoten zu der Artausschlüsselung. Es gilt nun folgendes: Wird auf dem Wege ein Knoten mit Positiv-Entscheidung zum Folgeknoten (auf dem Weg) verlassen, so treffen die in dem Knoten formulierten Merkmale auf die Art in der Artausschlüsselung zu. Die allgemeineren Merkmale, durch die die Art charakterisiert ist, sind also auf den ganzen Weg verteilt und haben daher einen erheblich größeren Umfang als die bei der eigentlichen Artausschlüsselung niedergelegte Merkmalmenge. Würde man also die Merkmale jeder Art vollständig auflisten, so würde sich ein um eine Größenordnung (geschätzt: ca. Faktor 8) höherer Gesamtinformationsumfang mit entsprechend höherem Speicherbedarf ergeben. (Bei einem vollständigen Binärbaum der Höhe  $h$  ergibt sich ein Kompaktifizierungsgewinn in der Größenordnung von Faktor  $h/2$ .) Nicht jedes Wissen ist für eine solche hierarchische Anordnung geeignet, dennoch scheint es lohnend zu sein, das Prinzip auch in anderen Anwendungen als bei der merkmalarientierten Objektbestimmung auf Anwendbarkeit zu prüfen.

Der dichotome Schlüssel nutzt die hierarchische Merkmalgliederung gezielt aus, so daß der Anwender sein Bestimmungsziel mit einem relativ geringen Spezifikationsaufwand erreicht. Kombinatorische Analysen des Moser-Schlüssels anhand der DTLEX-Datenbasis haben folgendes ergeben: Im Mittel werden für die Bestimmung einer Art auf jeweils kürzestem Wege 26,7 Entscheidungsschritte benötigt, bei Gattungen sind es 16,2 Schritte. Hierbei sind alle ca. 3100 Arten bzw. alle 173 Gattungen gleich gewichtet. Wie oben angedeutet wurde, ist der Schlüssel streng genommen kein Baum, sondern ein gerichteter Graph. Daher gibt es oft verschieden lange Bestimmungswege für denselben Zielknoten. Bestimmt man auf dem jeweils längsten Weg, so fallen im Mittel bei Arten 32,3 und bei Gattungen 20,2 Entscheidungsschritte an. Die Statistik muß daher zwischen den kürzesten und längsten Wegen unterscheiden. Im Sinne der mathematischen Informationstheorie bedeuten die oben genannten Werte folgendes: Eine Gattung kann im Mittel mit 16,2 Bit, eine Art mit 26,7 Bit Entscheidungsinformation identifiziert werden. (Das theoretisch nicht zu unterbietende Minimum wäre bei 3100 gleichwahrscheinlichen Arten  $\log_2(3100)=11,6$  Bit Entscheidungsinformation.)

#### 4. Gesamtübersetzung des Quelltextes in das Stammdatei-Format

Die bei Expertensystemen oft nicht übliche Trennung in Wissenserwerb und Wissenszugriff ist bei DTLEX sehr streng durchgeführt. Der gesamte Wissenserwerb erfolgt dadurch, daß der textuell vorliegende Bestimmungsschlüssel in die Struktur eines Entscheidungsbaumes übersetzt wird. Dies ist aber noch nicht ausreichend, um die gewünschte Funktionalität beim Bestimmen unbekannter Objekte zu realisieren, denn auch die synoptische Vorgehensweise soll unterstützt werden. Das Bestimmungssystem muß daher in der Lage sein, auch direkt vom Benutzer vorgegebene Merkmale in der gesamten Datenbasis auf Übereinstimmung/Nichtübereinstimmung prüfen und auswerten zu können, so daß sich der Benutzer auch ohne Entscheidungsbaum-Benutzung an das zunächst noch unbekannte Objekt herantasten kann. Allerdings ist ein Verfahren, bei dem nur Suchbegriffe und Schlüsselwörter im Text identifiziert werden, viel zu primitiv und ist bei unserer Anwendung von vornherein zum Scheitern verurteilt. Benötigt wird eine saubere, semantisch korrekte Extraktion der Merkmale aus den natürlichsprachigen Merkmalsbeschreibungen. Das hierfür entwickelte Verfahren wird im nächsten Abschnitt genauer beschrieben.

In diesem Abschnitt wollen wir bereits einen Schritt weiter gehen und das sogenannte **Stammdateiformat** von DTLEX vorstellen. Die Stammdatei ist ein isomorphes Abbild des oben eingeführten Entscheidungsbaumes. Dieser ist seinerseits ein mit global eindeutiger Adressierung ausgestatteter Gesamt-Bestimmungsschlüssel. Aus Systemsicht stellt die Stammdatei die gesamte Wissensbasis des DTLEX-Systems dar. Es handelt sich um eine relativ große Textdatei, in der jeder Knoten des Baumes in geeigneter textueller Form repräsentiert ist, angereichert um die Adressierung und Verweise und mit expliziter, semantisch annotierter Codierung der Merkmale des jeweiligen Knotens. Die Stammdatei ist noch gut lesbar, so daß man sich im Einzelfall davon überzeugen kann, was in ihr codiert ist.

##### Quelltextaufbereitung

Bevor die Gesamtübersetzung des Quelltextes angegangen werden konnte, mußte der Quelltext aufbereitet werden. Dabei wurden Überschriften mit dem Ende-Zeichen @ ausgestattet, um sie vom Folgetext eindeutig abzugrenzen. Am Ende jedes Teil-Bestimmungsschlüssels wurde das Endezeichen \$ eingefügt. Dadurch wurde es möglich, jeden Teilschlüssel im Gesamttext einfach zu erkennen und den gesamten Quelltext portionsweise zu übersetzen. Weiterhin wurde am Beginn jeder Zeile eine Seiten-Zeilen-Nummer eingefügt. Diese Zeilennummern sind dann normaler Bestandteil des Textes und haben sich als unerlässlich erwiesen, um bei Unklarheiten und Kontrollen eine schnelle 1-1-Beziehung zwischen Buch und Textdatei herzustellen. Um bei der Textanalyse Komplikationen zu vermeiden, mußten auch alle getrennten Wörter durch ihre ungetrennte Form ersetzt werden. Schließlich mußten auch alle Auslassungs-Bindestriche eliminiert werden. Z. B. wurden umgesetzt "hell- bis dunkelrot" in "hellrot bis dunkelrot", "Holzäste und -reste" in "Holzäste und Holzreste".

Fast alle diese Textaufbereitungen konnten mit schnell geschriebenen Programmen automatisch durchgeführt werden.

Der unten abgedruckte Ausschnitt aus der Moser-Stammdatei entspricht dem Teil des Bestimmungsschlüssels, der in Abschnitt 2 als Textbeispiel 2.2 wiedergegeben wurde, wobei aber aus Platzgründen ein Teil unterdrückt wurde:

#### **Textbeispiel 4.1: Ausschnitt aus der Stammdatei** (Quelltext siehe Textbeispiel 2.2)

---

§Adr:[2.2.10+0]  
§Entw:  
§Typ: Gattung  
@Überschrift: Moser-Überschrift: 2.2.10 Boletus Dill. ex Fr.,  
Röhrlinge  
@Gatt@Name: Gattung: Boletus#W  
@Autoren: Autoren: Dill. ex Fr.,  
@DGatt@Name: Gattungsname deutsch: Röhrlinge#W  
@Ko: {Röhrlinge im engeren Sinne}  
@Fl@Strktr: Frkp. fleischig#D#M  
@St@Frm: St. meist dickbauchig#B (Ausnahme 2.2.10.2.1)  
@St@Tex: mit feinflockiger#C oder netziger#C#X Oberfläche  
@H@Tex: H. sehr bald trocken#C  
@H@Tex: fein filzig#C oder kahl#C#X  
@Spp@Fa: Spp. oliv#Si bis olivbraun#Siq  
@Sp@Frm: Sp. spindelig#B oder fast spindelig#B  
@Sp@Tex: glatt#C#X  
@Rö@Tra@Hyph@Strktr@Fa@Frm: Rö.-Tr. mit stark divergierenden#B, locker#M verflochtenen#B, ±  
hyalinen#Sa#X  
oder blassen#S0 Hyphen#H#X in den Seitenschichten  
@Rö@Tra@Pig@Fa@N: Rö.-Tr. im Mediostratum satter gefärbt (wenn pigmentiert#K)  
@Pig: Pigmente#K bisher Variegatsäure#K  
@Pig: Variegatorubin#K und Xerocomsäure#K bekannt  
§Pvw:[2.2.10+1]  
§End:  
§Adr:[2.2.10+1]  
@Hym@Po@Fa: Hymenophor#X mit ± roten#Sn, rotorange#Snl, orange#Sl, orangegelben#Slk (und  
später etwas rot#Sn werdenden) Po.  
§Pvw:[2.2.10.1+0]  
§Nvw:[2.2.10-1]  
§End:  
§Adr:[2.2.10-1]  
@Hym@Po@Fa: Hymenophor#X mit weißen#Sb oder ± gelben#Sk Po.  
§Pvw:[2.2.10+2]  
§End:  
§Adr:[2.2.10+2]  
@Hym@Po@Fa: Hymenophor#X mit von Jugend an gelben#Sk Po.  
§Pvw:[2.2.10.2+0]  
§Nvw:[2.2.10-2]  
§End:  
§Adr:[2.2.10-2]  
@Hym@Fa: Hymenophor#X jung weißlich#Sb0  
@Hym@Fa: später erst gelb#Sk  
@Hym@Fa@Fbr: grünlichgelb#Sh0k verfärbend#U  
§Pvw:[2.2.10.3+0]  
§End:  
§Adr:[2.2.10.1+0]  
§Typ: Sektion  
§Entw:  
@Überschrift: Moser-Überschrift: 2.2.10.1 Sekt. Luridi Fr.  
@Po@Fa: Poren#X orange#Sl bis rot#Sn

§Pvw:[2.2.10.1+1]  
§End:  
§Adr:[2.2.10.1+1]  
@St@Tex@N: St. ohne Netz#C  
@St@Tex: kahl#C#X oder feinflockig#C oder schuppig#C  
§Pvw:[2.2.10.1+2]  
§Nvw:[2.2.10.1-1]  
§End:  
§Adr:[2.2.10.1-1]  
@St@Tex: St. mit deutlicher, netzartiger#C Zeichnung  
§Pvw:[2.2.10.1+5]  
§End:

.....**Auslassung Nummern 2. bis 4.**

§Adr:[2.2.10.1+5]  
@St@Frm@Tex: Netz#C grob u. langgezogen#B  
@St@Fa: St. meist ± orange#Sl  
@St@Bas@Fa: Basis weinrot#S2n  
@H@Fa: H. von olivbraun#Siq  
@H@Fa: orangebraun#Slq bis dunkel-braun#S2-q  
@H@Fa: auch rötlich#Sn0  
@FI@Fa: FI.#D blaß#S0  
@Bas@FI@Fa: in Basis oft weinrötlich#S2n0  
@FI@Fbr: schwach blauend#U  
@Po@Fa: Po. rotbraun#Snq  
@Sp@Mmü: Sp. 9-17/5-7 müm#Q  
@Bio: Laubwald#Tca u. Nadelwald#Tba  
@Abb: Kbch.#X 15, 16  
@Abb: MH#X I 158  
@Abb: Ha#X 4  
@Abb: MP#X 70  
@Abb: SPT#X II 59  
@Abb: R#X 129  
@Abb: Sing.#X VI 19, 18  
@Abb: C#X 275, 702  
@Abb: PD#X 62, 63  
@Abb: Md#X 160  
@Abb: LE#X 46  
§Entw:  
@Art@Name: Art-Name: Boletus+luridus#W  
@Autoren: Autoren: Schff.#X ex Fr.  
@DArt@Name: Artname deutsch: Netzstieliger+Hexenröhrling#W  
@MSeite: Moser-Seite: 66  
@MDkl: Moser-Dkl: 2.2.10.1.5  
@Eßbk: Eßbarkeit: (\*?);  
§Nvw:[2.2.10.1-5]  
§End:  
§Adr:[2.2.10.1-5]  
@St@Tex: Netzmaschen#C fein  
@St@Frm@N: nicht langgezogen#B  
§Pvw:[2.2.10.1+6]  
§End:

---

Anhand dieses Beispiels kann nun erläutert werden, was die Stammdatei im einzelnen enthält und was die diversen Kennungen bedeuten. Das Wichtigste in einer kurzen Zusammenfassung:  
- Strukturelemente mit dem Charakter von Wortsymbolen beginnen mit dem Zeichen §:  
§Adr:=Adresse des Entscheidungsbaum-Knotens, §Pvw:=Positiv-Verweis, §Nvw:=Negativ-Verweis, §End:=Ende des aktuell beschriebenen Knotens des Entscheidungsbaumes.

- Merkmaltypen beginnen mit @. Beispiele: @St=Stiel, @H=Hut, @Tex=Textur, @Frm=Form, @N=Negation, @Abb=Abbildungszitat. (Insgesamt gibt es in der DTLEX-Stammdatei 84 solche Merkmaltypen.)

- Komplexe Merkmaltypen werden durch Konkatenation der Einzelmerkmaltypen gebildet:  
@Bas@Fl@Fa:=Basis+Fleisch+Farbe.

- Merkmale werden grundsätzlich durch Text mit vorangestelltem Merkmaltyp aufgeschrieben.

- Der Merkmaltext, speziell einzelne Wörter darin, enthalten Annotationen, die unmittelbar an das Wort mit # gekennzeichnet folgen. Dabei bedeutet z. B.: #B=Formangabe, #Tca=Biotop, c=Laubwälder, a=Wälder, #W=Name, #X=Heißwort.

(Fachbegriffe, die nicht jedem geläufig sind, sowie unverständliche Abkürzungen wie z. B. bei den häufigen Abbildungszitaten sind als Heißworte (hot words) gekennzeichnet. Bei allen Textausgaben von DTLEX werden diese Begriffe zusätzlich gekennzeichnet, der Benutzer kann jederzeit, also bei jeder Eingabeaufforderung die Bedeutung vom System erfragen und erhält dann eine Erläuterung, worum es sich genau handelt.)

Aus dem obigen Ausschnitt aus der Stammdatei wird deutlich, daß auch Attribute wie "Eßbarkeit" und die Namengebung etc., praktisch alles, was im Moser-Quelltext steckt, in Merkmale umgesetzt wurde. Die Annotationen von Wörtern mit #... könnten auch dem Lexikon entnommen werden. Dies wäre aber bei der Durchmusterung großer Textbereiche sehr aufwendig und würde zu größeren Antwortzeiten führen. Es empfiehlt sich daher, alles was man über Worte und Merkmale weiß, in kompakter Annotation in den Text einzufügen, um Texte dann sehr schnell durchzumustern zu können. Bei der Textausgabe auf dem Benutzerbildschirm werden diese Kennzeichnungen selbstverständlich entfernt. Ein relativ einfaches Textfilter kann aus der Merkmalcodierung wieder den Original-Buchtext rekonstruieren. Heißworte behalten allerdings bei der Bildschirm-Ausgabe eine Kennzeichnung, damit der Benutzer sieht, daß er über sie zusätzliche Erklärungen erhalten kann.

Das Generieren der Stammdatei aus dem Moser-Quelltext ist ein relativ komplexer Vorgang, in dem vieles angewandt wird, was man aus dem Compilerbau kennt. Die wesentlichen Bestandteile des Übersetzungsprogrammes sind:

### **Lexikon-Modul**

Das Lexikon enthält zunächst alle Wörter, die im Quelltext, speziell in den Merkmalbeschreibungen vorkommen. Jedem Wort sind Attribute zugeordnet, die Beziehungen zu den Merkmaltypen herstellen. Im Lexikon sind aber auch die Merkmaltypen selbst und die Beziehungen zwischen denselben niedergelegt. Um schnelle Zugriffe und flexible Funktionalität sicherzustellen, wurde das Lexikon als AVL-Baum im Arbeitsspeicher aufgebaut. Extern ist es als Textdatei gespeichert und enthält für den Moser-Text ca. 10000 Einträge, allerdings ohne die ca. 3000 Art- und Gattungsnamen, die im Lexikon nicht benötigt werden, weil sie in der Regel durch ihre syntaktische Stellung klassifiziert werden.

### **Syntax-Modul**

Eine sinnvolle Zerlegung eines großen strukturierten Textes ist nur mit syntaktischen Analyseverfahren möglich. Das speziell für die DTLEX-Entwicklung konzipierte Verfahren wendet die Methode der rekursiven Suche oder besser des rekursiven Matchings an, allerdings mit einigen der Beschleunigung dienenden Verfeinerungen. Die Syntax selbst ist in gewöhnlicher Backus-Naur-Form in einer Textdatei abgelegt und hat ca. 200 Nichtterminalzeichen und ca. 550 Regeln bzw. Produktionen. Sie beschreibt nicht die Gesamtstruktur des Quelltextes in hierarchischer Form, wie das bei Programmiersprachen üblich ist, sondern nur die für die Struktursuche interessanten Teile. Dazu gehören z. B. Nummern, Verweise, Artausschlüsselungen, Überschriften, Autorennamen usw. Im Prinzip wird immer im noch nicht verarbeiteten Quelltext, also rechts von der aktuellen Arbeitsposition, nach bestimmten Mustern wie Satz, Nummern, Verweisen, Artausschlüsselungen usw. gesucht und dann der Text im Sinne der gefundenen syntaktischen Strukturen weiter zerlegt. Das Syntax-Modul hat eine einfache und flexible Schnittstelle. Jedes Nichtterminalsymbol kann als

Suchmuster verwendet werden. Das eigentliche Übersetzungsprogramm hat ohne die daran angeschlossenen Verarbeitungsvorgänge nur einen Umfang von ca. 300 Zeilen.

### **Zerlegung von Sätzen in Merkmale: Transformationsmodul**

Über diesen Programmteil wird im folgenden Abschnitt ausführlicher berichtet.

Die Gesamtübersetzung des Quelltextes von 1.4 MByte in die Stammdatei (2.8 MByte) ist ein relativ komplexer Vorgang und erfordert auf einem Apple Powerbook mit 20 MByte Hauptspeicher und 68LC040 Prozessor eine Rechenzeit von ca. 8 Stunden.

## **5. Automatische Merkmalextraktion aus natürlichsprachigem Quelltext**

### **Ziele der Einzelmerkmal-Erfassung**

Die Einzelmerkmal-Codierung wird an sich gar nicht benötigt, wenn der Entscheidungsbaum nur als dichotomer Bestimmungsschlüssel benutzt werden soll, denn dann werden dem Benutzer die Original-Merkmalbeschreibungen genau wie im Buch-Schlüssel angeboten. Soll der Benutzer jedoch die auf das zu bestimmende Objekt zutreffenden Merkmale im Sinne der synoptischen Methode vorgeben können, dann müssen die Merkmale in vergleichbare Klassen eingeteilt werden, um mit einem einfachen Matching-Verfahren Bewertungen vornehmen zu können. Wir können also festhalten: Das Ziel der Einzelmerkmal-Erfassung ist es, Merkmalklassen zu identifizieren, die dem Benutzer des Systems zur Detail-Spezifikation angeboten werden können. Es muß dem System dann möglich sein, zutreffende bzw. nicht zutreffende Merkmale unterscheiden zu können. Völlig in die Irre würde z. B. führen, wenn bei der systeminternen Merkmalbewertung Stiel-Farbangaben mit Hut-Farbangaben verwechselt werden würden.

### **Merkmal-Entschlüsselung aus natürlichsprachigen Texten**

Wie können Merkmale in natürlicher Sprache so aufbereitet werden, daß sie mit einfachen Techniken und Algorithmen gehandhabt werden können? Seit langem wird darüber geforscht, wie der Inhalt natürlichsprachiger Formulierungen maschinell erfaßt und verarbeitet werden kann. Z. B. erfordert die automatische Sprachübersetzung einen derartigen Arbeitsschritt. Dieser Fragenkomplex ist in zahllosen Veröffentlichungen über die Verarbeitung natürlicher Sprache bearbeitet worden. Bedeutende Teilerfolge wurden erzielt, es gibt auch für natürlichsprachige Mensch-Maschine-Schnittstellen schon recht brauchbare Systeme. Einen guten Überblick über Informatik-Methoden für die Verarbeitung natürlicher Sprache liefert z. B. das Buch von J. Allen [All88]. Ein speziell für die deutsche Sprache entwickelter Zugang zum maschinellen Sprachverständnis ist von M. Th. Rolland [Rol94] erarbeitet worden und hat uns manche Anregung gegeben.

Bei unserer Textvorlage haben wir es mit einer fast stenogrammartig komprimierten Beschreibung von Merkmalen zu tun. Zum Glück fehlen kompliziertere syntaktische Strukturen und vor allem höhere Abstraktionsebenen fast völlig, denn die Merkmalschilderung bezieht sich stets auf direkt zuzuordnende Eigenschaften. Nach einigen Versuchen zur Zerlegung der Sätze "von Hand" in brauchbare Merkmal-Codierungen wurde erkannt, daß ein relativ einfacher Umsetzungsmechanismus in fast allen Fällen das gewünschte Ergebnis, also eine Identifikation bzw. Klassifikation der Merkmale, liefert. Er wird jetzt genauer beschrieben.

### **Vorgehen bei der automatischen Merkmalextraktion**

Das in Abschnitt 4 bereits erwähnte Transformationsmodul, kurz "Trafo" genannt, hat die Aufgabe, natürlichsprachige Sätze in die Merkmalcodierung zu zerlegen. Die Umsetzung erfolgt in drei Phasen:

- (1) Zerlegung von Sätzen in Klauseln,
- (2) Bestimmung der Merkmalstypen, die zu jeder Klausel gehören,
- (3) Aufbereitung, Typzuordnung und Ausgabe der Merkmale.

Das Verfahren soll an einem Beispiel erläutert werden, das für unseren Text typisch ist:

H. mit (metallisch) grünlichem Ton, seidig-faserig, olivlich-graubräunlich, olivocker schimmernd, 1,5-3,5 cm, faserig bis fein schuppig, kegelig-glockig, ungerieft.

Der Satz wird zunächst in sogenannte Klauseln zerlegt:

H. mit (metallisch) grünlichem Ton,  
seidig-faserig,  
olivlich-graubräunlich,  
olivocker schimmernd,  
1,5-3,5 cm,  
faserig bis fein schuppig,  
kegelig-glockig,  
ungerieft.

In der Regel liefert die Kommasetzung bereits eine korrekte Zerlegung in Klauseln. Es gibt aber einige Ausnahmeregelungen, durch die in bestimmten Fällen auch an Stellen ohne Komma eine Zerlegung erfolgt oder ein Komma nicht als Klauseltrennstelle genommen wird. Im nächsten Schritt werden zu jeder Klausel die Merkmalstypen bestimmt, die in ihr vorkommen. Hierbei werden also zu allen Wörtern und Abkürzungen im Lexikon Attribute erfragt, speziell die Merkmalstypen. Das Resultat sieht so aus:

@Hut@Farbe@Farbton@Textur: H. mit (metallisch) grünlichem Ton  
@Textur: seidig-faserig  
@Farbe: olivlich-graubräunlich  
@Farbe@Textur: olivocker schimmernd  
@Maßcm: 1,5-3,5 cm  
@Textur: faserig bis fein schuppig  
@Form: kegelig-glockig  
@Textur: ungerieft

Im nächsten Schritt werden alle Klauseln daraufhin geprüft, ob sie einen Merkmalstyp haben, der als Satzgegenstand oder besser als Merkmalgegenstand zulässig ist. Merkmale müssen also immer einen Gegenstand enthalten, auf den sich die Eigenschaften beziehen. In obigem Beispiel hat nur die erste Klausel einen solchen Typ, nämlich @Hut. Dieser wird dann solange nach unten durchkopiert bzw. weitergereicht, bis alle Klauseln einen Merkmalgegenstand haben. Im allgemeinen werden bei diesem Prozeß Merkmalgegenstände auch rückwärts propagiert. Es wird ferner geprüft, ob sich unzulässige Typkombinationen ergeben. In manchen Fälle werden aus vorhandenen Typen neue zusätzlich generiert oder bei bestimmten Typkombinationen einzelne Typen ausgelöscht. Das für diese ganzen Vorgänge erforderliche Regelsystem ist im Lexikon gespeichert. Unzulässige Typisierungen von Klauseln werden als Fehler gemeldet und müssen durch Änderungen des Quelltextes behoben werden, insgesamt fielen dabei ca. 1000 Eingriffe in den 443 Seiten langen Quelltext an.

Zum Schluß werden noch bei einigen Wortklassen mit dem #-Mechanismus Wortannotationen angebracht und die Typen durch Kürzel ersetzt, also abgekürzt. In dieser Form wird der Satz schließlich in die Stammdatei übernommen:

@H@Fa@Fton@Tex: H. mit (metallisch#C) grünlichem#Sh0 Ton  
@H@Tex: seidig-faserig#C  
@H@Fa: olivlich-graubräunlich#Si0-dq0  
@H@Fa@Tex: olivocker#Sip schimmernd#C  
@H@Mcm: 1,5-3,5 cm#I  
@H@Tex: faserig#C bis fein schuppig#C  
@H@Frm: kegelig-glockig#B  
@H@Tex: ungerieft#C

In diesem Beispiel kommen an Wortannotationen vor: #C=Textur, #Sh0=Farbe+grün+lich, #Si0-dq0=Farbe+oliv+lich+--+grau+brown+lich, #I=cm-Maß, #B=Form. Speziell die Farben werden so codiert, daß auch Ähnlichkeitsvergleiche und sogar eine automatische Umsetzung in die aus der Computergraphik bekannte Rot-Grün-Blau-Norm möglich wäre. Derzeit verwendet DTLEX die Codierung nur, um dem Benutzer die Farben nach Ähnlichkeitsgruppen anbieten zu können, so daß er z. B. alle Grün-Töne in einem Schritt selektieren kann. (Insgesamt enthält der Moser über 2000 verschiedene Farbbezeichnungen, die allerdings hauptsächlich durch Kettenbildungen wie z. B. bei "olivlich-graubrünlich" entstehen.)

Die folgende Übersicht zeigt die Liste aller Merkmalstypen, die derzeit in DTLEX erfaßt werden:

*Eigenschaften:*

@Farbe, @Farbreaktion, @Farbton, @Fleisch, @Form, @Größe, @Grundfarbe, @Maß, @Maßcm, @Maßmm, @Maßmüm, @Merkmal, @Mitte, @Moser-Farbcode, @Reakt., @Struktur, @Text.-Gewebe, @Textur

*Hutbezogene Gegenstände:*

@Buckel, @Hut, @Rand, @Scheibe, @Scheitel

*Stielbezogene Gegenstände:*

@Gürtel, @Knolle, @Manschette, @Ring, @Spitze, @Stiel, @Stmaß, @Velum, @Volva

*Gewebebezogene Gegenstände*

@Haut, @Hyphen, @Mycel, @Pigmente, @Schnallen, @Trama, @Zystiden,

*Hymeniumbezogene Gegenstände*

@Basidien, @Fläche, @Gleba, @Hymenophor, @Lamellen, @Poren, @Röhren, @Schneide

*Sporenbezogene Merkmale:*

@Porus, @Sp.-Pulver, @Sporen

*Wachstumsbezogene Gegenstände:*

@Verbreitung, @Wuchsform, @Wuchszeit, @Biotop/Substrat

*Sonstiges:*

@Abbildung, @Figur, @Cortina, @Flagellen, @Frkp., @Genußwert, @Geruch, @Geschmack, @Habitus, @Kommentar, @Negation

Die oben vorgestellte automatische Merkmalextraktion funktioniert mit dem Moser-Text sehr gut. Mit den 9 Quelltext-Seiten 50, 100, 150, 200, 250, 300, 350, 400 u. 450 als Stichprobe wurde von Hand überprüft, wie genau die automatische Merkmalzuordnung funktioniert. Es wurden insgesamt 11 Fehler entdeckt in insgesamt 1438 Merkmalen, was einer mittleren Fehlerrate von 0.76 % entspricht. Wie solche Fehler zustandekommen, soll noch kurz an einem Beispiel erläutert werden. In den vier letzten Zeilen von Textbeispiel 4.1 (Stammdatei-Ausschnitt) ist ein solcher Fall zu sehen. Hier wird in die Stammdatei übernommen:

§Adr:[2.2.10.1-5]

@St@Tex: Netzmaschen#C fein

@St@Frm@N: nicht langgezogen#B

§Pvw:[2.2.10.1+6]

Zwar werden die Netzmaschen richtig als Stiel-Textur interpretiert, aber die Klausel "nicht langgezogen" wird nicht mehr auf die Textur, wie es richtig wäre, sondern auf den Stiel bezogen, der als Gegenstand von der vorhergehenden Klausel übernommen wird. Die über den Positiv-Verweis 2.2.10.1+6 erreichbaren Pilze haben zwar alle keinen langgezogenen Stiel, insofern ist das Merkmal zufällig korrekt. Das Merkmal ist auch negiert, daher wird es bei der Merkmalvorgabe durch den Benutzer generell nicht berücksichtigt, denn dabei haben sich die negierten Merkmale als außerordentlich problematisch herausgestellt, weil der Benutzer nur Positiv-Merkmale eingibt, woraus man nicht schließen darf, daß die nicht genannten Merkmale negativ zu werten sind. Bei Anwendung der Entscheidungsbaum-Methode tritt das Problem nicht auf, da hier der Text ohne Merkmalklassifikation, also stets so wie er lautet, zu beurteilen ist. Daher ist es zweckmäßig, die durch Merkmalmethode gewonnenen Bestimmungsdiagnosen evtl. durch normale Benutzung des Bestimmungsschlüssels zu überprüfen.



Eine noch präzisere, semantisch korrektere Erfassung der Merkmale kann nur durch eine umfangreichere Regelbasis erreicht werden, als wir sie bei obigem Verfahren einsetzen. Wichtig ist vor allem, daß alle Bezüge zwischen den Wörtern eines Satzes, ja sogar im Satz-Kontext, erfaßt werden; daß alle semantischen Eigenschaften der Wörter bekannt sind und dann aus dieser Analyse zunächst alle möglichen Interpretationen und schließlich die wahrscheinlichsten generiert werden. Es ist geplant, diesen Teil von DTLEX noch einmal zu überarbeiten und mit einem genaueren Analyseverfahren auszustatten. Derzeit ist auch das Zerlegen komplexerer Merkmalbeschreibungen in Einzelmerkmale noch sehr unbefriedigend. Das folgende Beispiel zeigt das Problem auf:

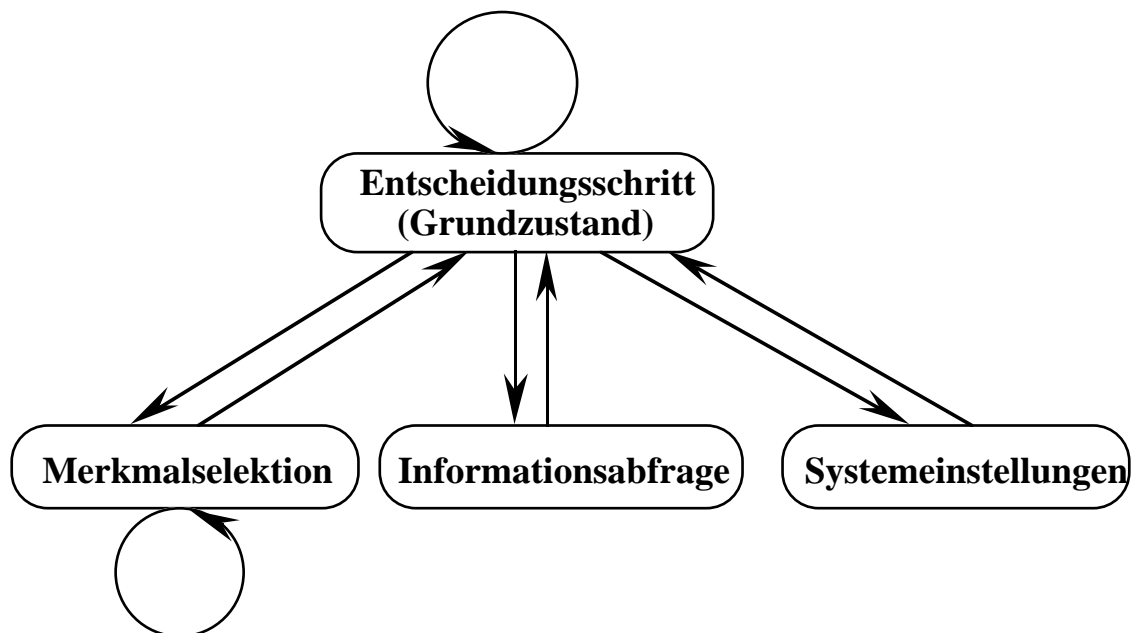
@Fl@Strktr: fleischige#D#M  
@H@Frm@Grö@Tex: kleine bis mittelgroße Arten mit radial#X faserigem#C bis rissigem#C  
oder wirr faserigem#C, häufig kegeligem#B aber manchmal auch konvexem#B H.

Die zweite Klausel ist bedingt durch die "mit"-Struktur sehr lang, denn "mit" ist an das Hauptwort H. = Hut gebunden. Dadurch wird die Merkmalklasse @H@Frm@Grö@Tex überladen, was für die synoptische Vorgehensweise nicht günstig ist. Es wird daher durch die Überarbeitung angestrebt, solche Klauseln in mehrere Einzelmerkmale zu zerlegen, z. B. so:

@Frkp@Grö: kleine bis mittelgroße Arten  
@H@Tex: radial#X faserig#C bis rissig#C  
@H@Tex: oder wirr faserig#C  
@H@Frm: häufig kegelig#B  
@H@Frm: manchmal auch konvex#B H.

## 6. Navigieren in Merkmal-Entscheidungsäumen

Angestrebt wird eine Symbiose zwischen der Schlüsselbearbeitung (Entscheidungsschritt im dichotomen Bestimmungsschlüssel) und der Merkmalvorgabe (synoptisches Verfahren). Der Benutzer soll in jeder Phase der Bestimmung zwischen beiden Vorgehensweisen hin und her schalten können. Außerdem soll er zu jedem Zeitpunkt Informationen abfragen können, z. B. über die bezüglich der Merkmalvorgaben am höchsten bewerteten Objekte. Die oberste Ebene der Dialogstruktur von DTLEX hat demgemäß folgendes einfaches Zustandsdiagramm:



Eine wichtige Grundlage für die weiteren Ausführungen ist der Begriff des **Suchraums**. Er ist definiert als die Menge aller Knoten des gesamten Entscheidungsbaumes, die zu einem bestimmten Zeitpunkt der Objektbestimmung noch nicht als abgearbeitet gelten. Bei Entscheidungsschritten wird der Suchraum stets kleiner, manchmal entfallen große Teile des Entscheidungsbaumes, die dann nicht mehr erreichbar sind.

#### Entscheidungsschritt:

Der Grundzustand wird nicht verlassen, wenn die jeweils dem Benutzer vorgelegten Schlüssel-Entscheidungen normal beantwortet werden, also mit A="Nichtsternteil zutreffend", B="Sternteil zutreffend" oder U="Unbestimmt". Die Antwort "Unbestimmt" führt dazu, daß beide Wege im Bestimmungsschlüssel verfolgt werden, also keiner der Folgeknoten ausgeschlossen wird. Durch A-B-U-Entscheidungen wird der Suchraum stets eingeschränkt, weil die nicht mehr erreichbaren Knoten des Entscheidungsbaumes herausfallen, auf jeden Fall auch der bearbeitete Knoten.

Beim Bestimmen mit dichotomen Bestimmungsschlüsseln ist es wichtig, daß die einzelne Nummer stets als ganzes betrachtet wird. Der Benutzer soll sorgfältig prüfen und abwägen, ob er sich für den Nichtstern-Teil oder den Sternteil entscheidet. Daher muß das System beide Textteile zusammen anbieten. (In der ersten Version von DTLEX wurde gemäß der Definition des Entscheidungsbaumes stets nur ein Knotentext zur Entscheidung vorgelegt, was sich aber nicht bewährt hat.)

Die Kopftexte von Teilbestimmungs-Schlüsseln sind ein Sonderfall, hier werden die allgemeinen Merkmale von Gattungen, Sektionen usw. dokumentiert. Wenn ein solcher Knoten abzuarbeiten ist, muß der Benutzer mit J oder N entscheiden, ob die meist allgemeiner gefaßten Merkmale auf die zu bestimmende Art zutreffen.

#### Merkmalselektion:

Innerhalb des Zustands "Merkmalselektion" können nacheinander mehrere Merkmale für den noch unbekanntem Pilz vorgegeben werden. Angeboten werden Standard-Merkmale wie Hut-Durchmesser, Stiel- und Sporenmaße. Daneben kann aber auch jede andere Merkmalkombination frei gewählt und bearbeitet werden. Der Suchraum kann durch Merkmalselektion grundsätzlich nicht eingeschränkt werden, aber die intern geführten Kurswerte für die einzelnen Arten werden verändert. In der Informationsabfrage kann man sich jederzeit die Arten nach absteigendem Kurswert sortiert anschauen.

#### Informationsabfrage:

Diesen Zustand steuert der Benutzer an, wenn er z. B. wissen möchte, welche Arten bezüglich des synoptischen Verfahrens den höchsten Kurswert haben, also gut zu den vorgegebenen Merkmalen passen. Die Arten werden nicht nur mit Namen, sondern mit dem vollen Merkmaltext der Artausschlüsselung im Moser gezeigt, was eine große Hilfe ist, um relevante und nicht relevante Arten schnell zu erkennen, speziell wenn nur wenige Merkmale eingegeben wurden. Es gibt hier auch einen sogenannten Lagebericht, der die 40 Arten mit den höchsten Kurswerten und die noch im Suchraum befindlichen Gattungen nur mit Namen aufzählt.

#### Systemeinstellungen:

Wichtig ist hier vor allem die Möglichkeit, den Suchraum bewußt einzuschränken, z. B. auf eine oder mehrere Gattungen. Denn oft kann der etwas versiertere Mykologe die Gattung einer Art problemlos angeben, ohne schon zu wissen, um welche Art es sich handelt.

Oben wurde festgestellt, daß der Suchraum allein durch Merkmalvorgaben nicht eingeschränkt werden kann. Ausgenutzt werden die Kurswerte jedoch von dem Mechanismus, der den nächsten Entscheidungsschritt bestimmt. Wenn mehrere U=Unbestimmt-Entscheidungen stattgefunden haben, so gibt es auch mehrere mögliche Knoten des Entscheidungsbaumes, die als nächstes bearbeitet werden können. Hierbei wird dann der Knoten bevorzugt, in dessen Folgeknotenbereich die Art mit höchstem Kurswert liegt. Diese wird zielstrebig angesteuert. Wenn also bereits hohe Kurswerte existieren, so wird man z. B. durch aufeinanderfolgende U-Entscheide auf dem schnellsten Weg zur Art mit höchstem Kurswert geführt. Insgesamt wird durch diese Strategie allerdings nur ein marginaler Gewinn erzielt, der sich nur dann sehr positiv auswirkt, wenn man einzelne

Schlüsselfragen mit U beantworten mußte und durch Merkmalvorgaben schon signifikante Kurswerte erzielt hat.

## 7. Merkmalvorgaben und Gestaltung der Benutzerschnittstelle

Die Benutzerschnittstelle von DTLEX wurde an die natürlichen Gegebenheiten der inneren Struktur des Systems angepaßt, aber so gestaltet, daß Kurz- und Langzeitgedächtnis des Anwenders möglichst wenig belastet werden. Außerdem sollte bei der Benutzung ein möglichst geringer Tippaufwand anfallen. Der Benutzer wird in der Regel geführt. Was er einzugeben hat bzw. eingeben kann, wird stets kurz mitgeteilt, so daß der Benutzer nie mit Unklarheiten zu kämpfen hat. DTLEX arbeitet in einem großen bildschirmfüllenden Textfenster, das im Scroll-Modus betrieben wird: Der alte Text wird nach oben bewegt und verläßt oben auch das Fenster, während der neue Text unten auf dem Bildschirm erscheint. Die Benutzereingaben werden stets am unteren Textende erwartet. Wenn Selektionsmenues erforderlich sind, werden sie stets wie im folgenden Beispiel (Dialog-Protokoll-Text) ausgegeben, in dem Benutzereingaben fett gedruckt sind:

### Textbeispiel 7.1

---

Grundzustand: @+7 | Schlüsselfrage: Treffen folgende Merkmale zu?

A:

Spp. weiß.

B:

Spp. gefärbt.

Eingabe A, B, U=unklar? (M=Merkmale| i=Infos | S=Sys-Einstellungen | E=Ende)

**i**

Informationsausgaben.....

0 -> Zurück zum Bestimmungsschlüssel<N?

1 -> Arten in Kurswert<O-Reihenfolge auflisten?

2 -> Lagebericht ausdrucken?

3 -> Artmerkmale/Gattungsmerkmale ausgeben?

4 -> Vollständige Artmerkmale/Gattungsmerkmale ausgeben?

Bitte Ziffer eingeben:

**0**

Im Suchraum<P noch 2969 Zielobjekte<R. Maximaler Kurswert<Q : 0.0000

---

Grundzustand: @+7 | Schlüsselfrage: Treffen folgende Merkmale zu?

A:

Spp. weiß.

B:

Spp. gefärbt.

Eingabe A, B, U=unklar? (M=Merkmale| i=Infos | S=Sys-Einstellungen | E=Ende)

**s**

Systemeinstellungen<S.....

0 -> Zurück zum Bestimmungsschlüssel<T?

1 -> Beschränktes Suchgebiet<U definieren, z. B. eine Gattung auswählen?

2 -> Suchgebiet<V durch Ausschluß von Gattungen verkleinern?

3 -> Heißworte-Anzeige<W unterdrücken?

4 -> Bildschirmdialog auf Protokolldatei aufzeichnen?

5 -> Testmodus einschalten?

6 -> Textdarstellung im Intern-Format<X ? (Nur für Informatik-Gurus genießbar)

Ziffer eingeben

**0**

Im Suchraum<Y noch 2969 Zielobjekte<A. Maximaler Kurswert<Z : 0.0000

---

Grundzustand: @+7 | Schlüsselfrage: Treffen folgende Merkmale zu?

A:

Spp. weiß.

B:

Spp. gefärbt.

Eingabe A, B, U=unklar? (M=Merkmale| i=Infos | S=Sys-Einstellungen | E=Ende)

**m**

Merkmale des unbekanntes Pilzes eingeben.....

0 -> Zurück zum Bestimmungsschlüssel<B?

1 -> Hutdurchmesser<C in cm spezifizieren?

2 -> Stielgröße spezifizieren?

3 -> Sporengöße spezifizieren?

4 -> Merkmalkombinationen<D bearbeiten?

Nur eine der Ziffern eingeben.

**0**

-----

Im Grundzustand werden die Eingaben A, B, U="Unklar", M="Merkmale eingeben", i="Informationen nachfragen", S="Systemeinstellungen" und E="Ende der Bestimmung" zur Auswahl angeboten. Echte Menüs mit Erläuterungen, oben erscheinen die Menüs für "Informationsausgaben anstoßen", "Systemeinstellungen" und "Merkmale des unbekanntes Pilzes eingeben", werden in Zeilendarstellung präsentiert und durch Zifferneingabe selektiert. Der Bedienungsaufwand ist daher relativ gering, soll aber zukünftig durch optionales Anklicken der Benutzereingabe mit der Maus noch verringert werden. Die optionale Mausbenutzung hat den großen Vorteil, daß der Benutzer die Augen auf den Bildschirm richten kann und nicht immer wieder auch die Tastatur anschauen muß. In der Textausgabe werden alle Heißworte mit einer Annotation "<A", "<B" usw. bis "<Z" gekennzeichnet. Der Benutzer kann jederzeit mit Eingabe von z. B. "<C" Informationen über das so gekennzeichnete Heißwort erfragen. Wer die Heißwort-Annotation im Text nicht wünscht, kann diese im Menü "Systemeinstellungen" ausschalten.

Das Konzept der möglichst umfassenden Informationsdarstellung bringt es mit sich, daß häufig große Textmengen auf dem Bildschirm auszugeben sind. Daher muß die Zeilenlänge möglichst gut ausgenutzt werden. Mit Hervorhebungen durch zwei verschiedene Schriftgrößen sowie Fettschreiben und tabellarischem Ausrichten soll zukünftig eine für das Lesen gefälligere Textdarstellung erreicht werden.

Wir wollen uns nun dem etwas schwierigeren Problem der freien Merkmalvorgabe durch den Benutzer zuwenden. Das implementierte Verfahren verläuft in folgenden Schritten:

1. Dem Benutzer werden die im Suchraum auffindbaren Merkmalklassen angeboten.
2. Der Benutzer wählt eine Merkmalklasse aus.
3. Das System durchsucht die gesamte Datenbasis nach dieser Merkmalklasse und legt eine Liste aller gefundenen Merkmale an.
4. Die Merkmale können in verschiedenen Textdarstellungen auf dem Bildschirm aufgelistet werden und, falls vom Benutzer gewünscht, selektiert werden.
5. Das System erhöht den Kurswert der Arten, auf die die selektierten Merkmale zutreffen.

**Zu 1.:** Einige Merkmalklassen kommen in der Datenbasis sehr häufig, andere seltener vor. Der folgende Dialogausschnitt (Bildschirmausgabe + Benutzereingabe) liefert die Merkmalstatistik, wobei der Suchraum noch nicht eingeschränkt ist:

---

Die häufigsten Merkmalkombinationen<L im Suchraum<M sind: ([...]=Anzahl des Vorkommens)

- |                              |                          |                            |
|------------------------------|--------------------------|----------------------------|
| 1 Abbildung[6062]            | 2 Farbe+Hut[4930]        | 3 Biotop[4083]             |
| 4 Autoren[3236]              | 5 Artname+Name[3063]     | 6 Moser-Seite[3063]        |
| 7 Moser-Dkl[3063]            | 8 Maßmüm+Sporen[3013]    | 9 Farbe+Lamellen[3010]     |
| 10 Hut+Maßcm[2715]           | 11 Farbe+Stiel[2496]     | 12 Form+Hut[1855]          |
| 13 Hut+Textur[1831]          | 14 Stiel+Stielmaß[1532]  | 15 Form+Lamellen[871]      |
| 16 Geruch[771]               | 17 Form+Sporen[723]      | 18 Stiel+Textur[678]       |
| 19 Artname-deutsch+Name[658] | 20 Farbe+Hut+Textur[641] | 21 Eßbarkeit[532]          |
| 22 Form+Stiel[482]           | 23 Sporen+Textur[446]    | 24 Farbe+Fruchtkörper[441] |

25 Überschrift[422] 26 Form+Zystiden[346] 27 Moser-Figur[328]  
28 Fruchtkörper+Größe[305] 29 Farbe+Hut+Rand[303] 30 Stiel+Struktur[295]  
31 Farbe+Fleisch[291] 32 Farbe+Hut+Stiel[285] 33 Hut+Rand+Textur[284]  
34 Geschmack[238] 35 Farbe+Stiel+Textur[223] 36 Basis+Farbe+Stiel[210]  
37 Verbreitung[209] 38 Farbe+Sporenpulver[190] 39 Farbe+Hut+Mitte[184]  
40 Farbe+Farbton+Hut[180]

Soll eine Merkmalkombination <N bearbeitet werden?

Falls gewünscht, bitte Nummer der Merkmalskombination eingeben. (Zwischen 1 und 40)

J=Auflistung fortsetzen | N=Beenden

**J**

41 Farbe+Lamellen+Schneide[179]  
42 Farbe+Spitze+Stiel[177] 43 Gattung+Name[173] 44 Kommentar[158]  
45 Basis+Form+Stiel[149] 46 Zystiden[145] 47 Form+Hut+Rand[136]  
48 Form+Lamellen+Struktur[134] 49 Hut+Negation+Textur[132]  
50 Farbe+Hut+Scheitel[132] 51 Abbildung+Moser-Figur[125]  
52 Farbe+Hut+Texturgewebe[125] 53 Maßmüm+Zystiden[125]  
54 Gattungsname-deutsch+Name[123] 55 Farbe+Velum[120]  
56 Lamellen+Struktur[113] 57 Hut+Struktur[113] 58 Stiel[109]  
59 Wuchszeit[103] 60 Fleisch+Struktur[101] 61 Hut+Reaktion[98]  
62 Farbe+Moser-Farbcode+Sporenpulver[97] 63 Farbe+Poren[96]  
64 Basidien[96] 65 Reaktion+Sporen[93] 66 Farbreaktion+Fleisch[91]  
67 Farbe+Fleisch+Reaktion[90] 68 Geruch+Negation[89]  
69 Form+Hut+Maßcm[89] 70 Delete[87] 71 Farbe+Farbreaktion+Hut[87]  
72 Fruchtkörper+Form[86] 73 Moser-Farbcode+Sporenpulver[84]  
74 Spitze+Stiel+Textur[84] 75 Farbe+Farbton+Lamellen[83]  
76 Farbe+Sporen[82] 77 Farbe+Hut+Scheibe[82]  
78 Farbreaktion+Stiel[81] 79 Größe+Sporen[76] 80 Biotop+Negation[75]  
81 Negation+Zystiden[74] 82 Habitus[73] 83 Merkmal[69]  
84 Negation+Reaktion+Sporen[69] 85 Größe+Hut[69] 86 Ring+Stiel[66]  
87 Maßcm+Stiel[64] 88 Fruchtkörper+Textur[63] 89 Farbreaktion+Hut[63]  
90 Wuchsform[62]

Soll eine Merkmalkombination <O bearbeitet werden?

Falls gewünscht, bitte Nummer der Merkmalskombination eingeben. (Zwischen 41 und 90)

J=Auflistung fortsetzen | N=Beenden

**N**

---

Den ersten Zeilen der Auflistung kann man z. B. entnehmen, daß das Merkmal Farbe+Hut[4930] also insgesamt 4930 mal in der Datenbasis von DTLEX vorkommt. Das Merkmal Wuchsform[62] auf Position 90 der Häufigkeitsrangliste kommt allerdings nur noch 62 mal in der Datenbasis vor.

Insgesamt enthält die Datenbasis rd. 65000 Merkmale. Es gibt über 2000 verschiedene Merkmalklassen, die durch überladene Kombinationen wie z. B.

@Rö@Tra@Hyph@Strktr@Fa@Frm zustande kommen. Solche Merkmale kann der Benutzer z. B. erfassen, indem er nur @Tra@Hyph als Merkmalklasse auswählt und diese Einzelmerkmale als Untermenge interpretieren läßt. Dann erhält er alle Merkmale, in denen sowohl @Tra=Trama als auch @Hyph=Hyphen vorkommt. Der Dialog bei der Merkmalvorgabe ist so gestaltet, daß der Benutzer solche Eingaben nie textuell vornimmt, sondern stets Elemente aus größeren oder kleineren Listen über Zahleingaben bestimmt oder in der geplanten Weiterentwicklung von DTLEX mit der Maus anklickt. Insbesondere für weniger geübte Computeranwender erhöht dies die Benutzerfreundlichkeit erheblich.

**Zu 2.:** Der Benutzer kann die Merkmalklasse entweder aus angebotenen Auflistungen auswählen, er kann aber auch frei gewählte Kombinationen vorgeben und muß sich entscheiden, ob er die Obermengen-Interpretation, die Exakt-Interpretation oder die oben erwähnte Untermengen-Interpretation wünscht.

**Zu 3.:** Das System durchsucht nur die Knoten des Suchraumes nach den vom Benutzer eingegebenen Merkmalklassen. Daher ist es meist sinnvoll, den Suchraum durch

Entscheidungsschritte erst einmal deutlich einzuschränken, denn dann werden wesentlich kleinere Merkmalmengen angeboten, was auch die Selektion vereinfacht.

**Zu 4.:** Bei der eigentlichen Einzelmerkmal-Selektion gibt es drei verschiedene Grundverfahren, zwischen denen sich der Benutzer entscheiden kann:

a) Maßeingaben

Maß-Merkmale wie Hutdurchmesser, Stiel- und Sporenmaße werden in speziellen Routinen abgehandelt, der Benutzer gibt die für den unbekanntten Pilz festgestellten Mittelwerte der einzelnen Maße zahlenmäßig ein. Mit einer speziellen Metrik wird die relative Abweichung der vorgegebenen Zahlenwerte von den gespeicherten Zahlenwerten berechnet und in einen individuellen Kurswert-Aufschlag zwischen 0.0 und 2.0 umgerechnet.

b) Selekt-Merkmale

Farben, Formen, Texturen, Biotop/Substrate usw. sind als Eigenschaftsangaben oder Selekt-Merkmale ausgezeichnet. Es sind dies genau die in der Stammdatei mit angehängtem #.. speziell annotierten Wörter. Diese Wörter werden aus dem Merkmaltext extrahiert und dem Benutzer zur Auswahl vorgelegt. So generierte Auswahllisten sind in der Regel in ihrem Umfang noch überschaubar und können, wenn auf dem Bildschirm in Häufigkeits- oder Alphabet-Sortierung präsentiert, noch mit erträglichem Aufwand ausgewählt werden. Die folgende Auflistung zeigt z. B. die für die Merkmalklasse Sporen+Form gefundenen Formen, eingeschränkt auf die sehr artenreiche Gattung Cortinarius. In eckigen Klammern [...] wird jeweils die Häufigkeit des Auftretens des einzelnen Begriffes mitgeliefert:

---

Ausgabe aller gefundenen Begriffe alphabetisch sortiert:

apfelkernförmig[1] birnenförmig[1] bohnenförmig[1] breit[9] eiförmig-ellipsoidisch[1]  
eiförmig-walzenförmig[1] ellipsoidisch[19] ellipsoidisch-mandelförmig[1] elliptisch[11]  
elliptisch-mandelförmig[1] kernförmig[1] kurz-elliptisch[1] länglich-elliptisch[1] mandelförmig[28]  
mandelförmig-sublimoniform[1] Plage[1] rundlich[35] schiffchenförmig[1] schlank[1] schmal[3]  
schmal-spindelrig[1] spindelförmig[2] spindelrig[6] spindelrig-schmal[1] sublimoniform[1]  
subzylindrisch[2] verjüngt[1] walzenförmig-zylindrisch[1] zitronenförmig[16]  
zylindrisch-walzenförmig[1]

Soll eine Selektion nach Relevanz vorgenommen werden? (J/N)

---

Die eigentliche Selektion erfolgt dann, indem jeweils 5 Begriffe gezeigt werden, wobei sich der Benutzer mit + oder - oder "Weiter" im einzelnen festlegen muß. Bei zwei Typen von Selekt-Merkmalen, nämlich bei Farbe und Biotop/Substrat wurde bereits in der Wortannotation eine Sub-Typisierung durchgeführt, so daß der Benutzer eine ganze Merkmalklasse wie z. B. "Holzsubstrat" oder "Braun-Töne" selektieren kann. Er hat aber auch hier die Möglichkeit, in Details einzusteigen und den einzelnen Merkmalbegriff zu selektieren.

c) Merkmaltext-Einzelprüfung

Dieses Verfahren wendet der Benutzer an, wenn er die Merkmalformulierungen einer Merkmalklasse im ganzen Text prüfen und daraufhin selektieren will. Das ist immer dann notwendig, wenn kein Selekt-Merkmal wie z. B. Farbe in der Merkmalklasse vorkommt oder wenn mehrere Selektmerkmale, z. B. Textur+Farbe im Zusammenhang geprüft werden sollen. Da die im Suchraum gefundenen Merkmale in Listen ausgegeben werden, kann man mit diesem Mechanismus auch ganz andere Informationen gewinnen, z. B. eine vollständige Liste aller Art-Namen im Suchraum.

Derzeit beruht also die Merkmalselektion von DTLEX grundsätzlich auf dem Prinzip, daß der Benutzer auswählen kann unter den in der Datenbasis gefundenen Merkmal-Begriffen. Für die Zukunft wäre es wünschenswert, auch eigene merkmalsbeschreibende Begriffe eingeben zu können. Das würde eine Ähnlichkeitsprüfung zwischen den eingegebenen Begriffen und den systeminternen Merkmalcodierungen voraussetzen, was zwar machbar erscheint, aber doch einigen zusätzlichen Implementierungsaufwand erfordern wird.

**Zu 5.:** Hat der Benutzer Merkmale selektiert, die das zu bestimmende Objekt hat, so erhöht das System den Kurswert der zu den selektierten Merkmalen gehörenden Arten, normalerweise um den Wert 1.0, bei Maßvergleichen um einen Abstandswert aus dem Intervall [0.0...2.0]. Dieser Vorgang läuft so ab: Ist das positiv zu wertende Merkmal in einem Knoten des Entscheidungsbaums angesiedelt, der einer Art ausschließung entspricht, so ergeben sich keinerlei Komplikationen, genau diese Art erhält den erhöhten Kurswert. Handelt es sich aber um einen Knoten, der keiner Art ausschließung entspricht, so liegen die Dinge etwas komplizierter, denn die Merkmale des Knotens treffen für alle Arten zu, die über den Positiv-Verweis des Knotens erreichbar sind. Daher müssen alle diese Arten jeweils die entsprechende Kurswert-Erhöhung erhalten. Das jetzt implementierte Kurswert-Verfahren beachtet nur positive, keine eindeutig negativen Merkmalübereinstimmungen. Zukünftig muß aber darüber nachgedacht werden, wie auch eindeutig negative Merkmalübereinstimmungen in einen Kurswert-Abschlag umgesetzt werden können.

## 8. Interne Programmstruktur von DTLEX

Die gesamte DTLEX-Software besteht aus drei größeren Programmen:

1. Übersetzer Quelltext -> Stammdatei (ca. 120 KB Code)
2. Umsetzung Stammdatei -> Ladedateien (ca. 50 KB Code)
3. DTLEX-Dialogprogramm (ca. 100 KB Code)

Die Umwandlung des dichotomen Bestimmungsschlüssels in das Stammdateiformat ist bereits erläutert worden. Der Schritt "2. Umsetzung..." überführt das Stammdatei-Format in mehrere interne Dateien, die dann ein schnelles Hochfahren und Starten des eigentlichen DTLEX-Dialogprogrammes erlauben. Dabei werden insbesondere alle Adressen und Verweise innerhalb des Entscheidungsbaumes auf Integer-Werte abgebildet.

Die eigentliche Expertensystem-Shell, also das DTLEX-Dialogprogramm, ist so realisiert worden, daß sich ein sehr günstiges Antwortzeitverhalten ergibt.

Die wesentlichen Datenstrukturen sind:

- a) Skelett des Entscheidungsbaumes (ohne Merkmale)  
Realisierung: Array [1..10000] of Record....
- b) Merkmal-Typenliste  
Realisierung: Array [1..70000] of Record....
- c) Merkmal-Textdatei auf Plattenspeicher  
Realisierung: File of String[40]

Die bezüglich des Antwortzeitverhaltens kritischste Operation ist die Suche und Durchmusterung aller Merkmaltexte nach einer bestimmten Merkmal-Typklasse dar. Aus der Merkmal-Typenliste (b) kann man sehr schnell die relevanten Merkmalnummern und ihre Position auf der Merkmal-Textdatei (c) gewinnen, denn die Typkombinationen sind als Set-Typ, also intern als Bit-Muster, dargestellt. Diese Datenstruktur entspricht im Prinzip einer invertierten Datei, mit der sichergestellt wird, daß genau nur die relevanten Merkmaltexte auf der Datei gelesen werden müssen. Wartezeiten entstehen praktisch nur noch bei den Dateizugriffen. Bei nicht eingeschränktem Suchraum erfordert der sehr häufige Merkmaltyp Hut+Farbe insgesamt mehr als 4000 Dateizugriffe, wobei dann 1150 verschiedene Farbangaben gefunden werden. Dieser Vorgang, es handelt sich praktisch um den "worst case", dauert auf dem Apple Powerbook 520 rd. 20 Sekunden, was in Anbetracht der dem Benutzer dann gelieferten Informationsmenge vertretbar erscheint: Die Ausgabe aller Farben umfaßt 215 Textzeilen, also einige Bildschirmseiten. Seltener vorkommende Merkmaltypen werden entsprechend schneller aufgesammelt, insbesondere bei bereits eingeschränktem Suchraum geht alles sehr schnell, weil selten mehr als 200 Merkmale zu erfassen sind. Alle anderen bei DTLEX intern anfallenden Operationen wie z. B. die Kurswert-Erhöhlungen können unter 1 Sekunde abgewickelt werden, weil kaum Dateizugriffe notwendig sind.

Das Programm benötigt ca. 2,5 MB Arbeitsspeicher, woraus man folgern kann, daß auch noch wesentlich umfangreichere Bestimmungsschlüssel bewältigt werden könnten: Bei 20 MB verfügbarem Hauptspeicher also rund 24000 Arten, was große Teile der Flora und Fauna Europas umfassen könnte.

Im nachhinein hat sich die frühe Entscheidung für eine maßgeschneiderte Implementierung in einer normalen Programmiersprache wie Pascal und nicht etwa PROLOG oder LISP als sehr richtig erwiesen, insbesondere unter dem Gesichtspunkt des Antwortzeitenverhaltens. Zwar arbeitet speziell das Syntaxanalyse-Verfahren bei der Erzeugung der Stammdatei sehr ähnlich zum rekursiven Matching-Verfahren von PROLOG, es ist aber nicht durch Unifikation und Listenhandhabung belastet und dadurch vermutlich unvergleichlich schneller. Auch die Portabilität eines gewöhnlichen Pascal-Programmes auf verschiedene Rechner-Plattformen ist erheblich günstiger zu beurteilen.

## 9. Schlußdiskussion

DTLEX zeigt einen neuen Weg zur Realisierung von Diagnosesystemen bzw. Systemen zu Objektbestimmung. Das System realisiert eine **Symbiose** von zwei unterschiedlichen Prinzipien, nämlich der Entscheidungsbaum-Methode und der klassischen Information-Retrieval-Methode. Jede dieser Vorgehensweisen hat ihre Stärken und Schwächen, durch das Zusammenspiel wird jedoch in jedem Fall eine bessere Lösung als durch eine Methode allein erzielt. Daraus ergibt sich noch eine weitere interessante Konsequenz: DTLEX könnte eine große Zahl von Objekten, die in flacher Struktur vorliegen, also nicht in einem Entscheidungsbaum eingegliedert sind, zur Bestimmung anbieten, wobei dann nur die synoptische Bestimmungsmethode zur Verfügung stehen würde.

Der **Wissenserwerb** durch automatisiertes Einlesen und Aufbereiten eines größeren Quelltextes, in unserem Fall des Moser-Bestimmungsschlüssels, bringt ganz erhebliche Vorteile gegenüber sonst bei Expertensystemen oft angewandten Methoden. So ist es bei DTLEX gelungen, ein System mit **umfassendem Inhalt** zu erstellen, was durch Von-Hand-Erfassen dieser Merkmalmenge nur mit einem vielfach höheren Aufwand möglich gewesen wäre.

DTLEX kann auch als ein Beitrag zur besseren Strukturierung zukünftiger **Multi-Media-Produkte** aufgefaßt werden. Schlagworte wie "lost in Hyperspace" deuten ja das Problem an, in großen Mengen von unstrukturiert bereitgestellten Informationsmengen semantisch gesteuerte Zugriffe vornehmen zu können. Die übliche Link-Technik von Hypertexten kann hier nicht das letzte Wort sein, vergleiche z. B. auch die Arbeit [BGP95]. Der häufigste Merkmaltyp in DTLEX ist @Abbildung[6062], es gibt also im Moser nicht weniger als 6062 Abbildungszitate. Wie schön wäre es, diese Abbildungen von Pilzen sofort auf dem Bildschirm anschauen zu können. Da das DTLEX-Stammdateiformat bezüglich neuer Merkmaltypen völlig offen ist, ließen sich alle möglichen Verweise selbst bis hin zu WWW-Adressen sofort und problemlos einbauen.

Die inhaltliche Erschließung von Merkmalbeschreibungen **in natürlicher Sprache** wurde bei DTLEX ebenso wie das ganze System im Stile der schnellen Prototyp-Entwicklung realisiert. DTLEX ist bei dieser Aufgabe nur deshalb relativ erfolgreich, weil die Sprache der Merkmalbeschreibung eine relativ einfache Struktur hat. Bei der Entwicklung wurden aber viele Einsichten gewonnen, speziell in die Natur und Struktur der natürlichen Sprache. Trotz enormer Forschungsanstrengungen, insbesondere auch in den USA, herrscht die Meinung vor, daß der Computer die natürliche Sprache noch nicht semantisch korrekt entschlüsseln kann. Die natürliche Sprache wirft also Probleme auf, die die Informatik noch nicht befriedigend gelöst hat, die aber, und das ist unsere feste Überzeugung, zukünftig gelöst werden müssen. Einige Anwendungen wie automatisches Abstracting werden seit langem versucht und es gibt sicher manche Teilerfolge. Es gibt aber noch viel mehr sinnvolle Anwendungen. Wann wird es soweit sein, daß der Computer für mich Bücher und Zeitschriften liest und mir dann im Dialog verdeutlicht, was mich am Inhalt des Buches evtl. interessieren könnte? Auch die automatische Übersetzung zwischen verschiedenen natürlichen Sprachen ist zu erwähnen. Wann sind spracherkennende Systeme kompetent in dem Sinn, daß sie wirklich genau verstehen, was ich in das Mikrofon gesprochen habe? Es besteht also zwischen der Welt der Computer und speziell dem Medium der natürlichen Sprache noch eine deutliche Kluft, die es Schritt für Schritt zu schließen gilt.



## Literatur

- [All88] James Allen  
Natural Language Understanding.  
The Benjamin/Cummings Publ. Co., Menlo Park, Cal. 1988.
- [BGP95] Basili R., F. Grisoli, M. T. Pazienza  
HIRMA: Hypertextual Information Retrieval System Managed by ARISTO.  
Data & Knowledge Engineering 17 (1995) 187-213.
- [Dob88] Dobbitsch P.  
Über Möglichkeiten des Computereinsatzes in der Mykologie.  
Z. f. Mykologie 54(1988), S. 113-118.
- [Kou93] Koune J. P.  
11. Europäische Cortinarientagung in Pays du Retz Loire, Atlantique F. vom 25.-31.  
Oktober 1993.  
Zeitschrift für Mycologie 60 (1994) S. 333.
- [MG90] Mendt A., M. Georgi  
Pilzbestimmung mit Hilfe der Maximum-Detektion.  
Z. f. Mykologie 56(1990), S. 187-198.
- [Mos83] Moser M.  
Die Röhrlinge und Blätterpilze.  
5. Aufl., Gustav Fischer Verlag, Stuttgart 1983.
- [Nei88] Neis B.  
PIDIAS - Entwicklung und Evaluation eines Expertensystems zur Bestimmung von Pilzen.  
Diplomarbeit, Universität Kaiserslautern, Fachbereich Informatik 1988.
- [Pup86] Puppe F.  
Assoziatives diagnostisches Problemlösen mit der Expertensystem-Shell MED2  
Dissertation, Fakultät für Informatik, Universität Kaiserslautern 1986.
- [Pup88] Puppe F.  
Einführung in Expertensysteme. Springer Verlag, Berlin 1988.
- [Rol94] M. Th. Rolland  
Sprachverarbeitung durch Logotechnik. Sprachtheorie, Methodik, Anwendungen.  
Dümmler-Verlag 1994.
- [Sch95] Schilling A.  
Software SYNOPSE - Ein Programm zur Pilzbestimmung.
- [SG83] Salton G., M. J. McGill  
Introduction to Modern Information Retrieval.  
McGraw-Hill, New York 1983.