McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

Applied Cladistics: New Models for Classification and Taxonomy Research; or How the *New York Review of Books* Taught Me Everything I Needed to Know about Taxonomy Research

Arthur McCaffrey
Digital Equipment, Corp.
129 Parker St.
Maynard MA 01754

This paper is about knowledge engineering and the design and development of knowledge representation structures to represent and support classification schemes in computers, particularly those based on the semantics of taxonomies and thesauri. The paper provides an interdisciplinary synthesis of research and methodology from many different sources-- artificial intelligence and knowledge engineering, computational linguistics, information retrieval-- but also from the new field of cladistics which has its origins in zoology, biology and paleontology.

This paper has the intent of applying cladistics models and taxonomies to the pragmatic problem of building computer applications for very intelligent and very powerful information retrieval.

My formulation of "applied cladistics" draws upon the modern work of Hennig, and its extensions by Sokal and Sneath, by Humphries and Parenti, by Platnick, and others. cladistics and numerical taxonomy between them represent two different approaches in recent years to building a better model and rationale for classification and taxonomics. Numerical taxonomy has been implemented via quantitative computer programs, whereas cladistics involves analysis and classification components that are more substantive and semantic. The subject matter of both approaches, however, has been natural history, evolution and paleontology. I will attempt to generalize these theoretical approaches to other disciplines.

This paper will further attempt to bring to bear on these recent development in cladistics research some insights from linguistics concerning classification "universals" and the notion of natural or "unmarked forms" of distinctive features or characteristics in systems (or what cladistics might refer to as "primitive vs. derived" characteristics). It is these latter notions of naturalness and recency of forms in evolution that provide a classification heuristic for other disciplines (such as Information Science) which can be used to architect powerful knowledge representation schema that are hierarchically structured and ordered, i.e., a "system".

To continue the inter-disciplinary flavor of the paper, it will be further argued that these insights from linguistics and modern cladistics taxonomic research can be combined with recent work on "super-thesauri" (Stern), and facet methodology for thesaurus construction (Rockmore) to provide a usable blueprint for building knowledge representation structures into a computer which will be capable of very sophisticated processing of textual information.

While not completely isomorphic with human thought processes, such data structures will have much in common with taxonomies and thesauri which are the formal articulation of human thought processes. Applied cladistics, therefore, is itself a metaphor for combining and synthesizing research from many disciplines in order to focus in a new way on the common but universal

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

problem of building knowledge and intelligence about the world into computers. Cladistics, taxonomics, and thesauri are ways that humans think about, reason about and classify the world--the new formalisms discussed in this paper should bring us one step closer to realizing that kind of knowledge and information in a machine.

APPLIED CLADISTICS

The three components of the history of life are form, time and space. The biological disciplines dealing with these are systematics, which concerns the variety of form, paleontology, concerned with that variety in time, and biogeography, concerned with that variety in space. cladistics calls into question traditional attitudes in all three, and offers a new approach to comparative biology which has a coherent theoretical base that is not necessarily tied to evolutionary theory. As a science of pattern, cladistics holds out the possibility of a reconstruction of the history of life in space and time that does not depend on Darwinian or neo-Darwinian presuppositions. The interest of that reconstruction or cladogram is that theories of process--neo-Darwinism or any other -- can be tested only against nature, and their best test will be their success in explaining past and present configurations of life. But if we are taught, as we have been, to see that pattern through the spectacles of evolutionary theory, how could the pattern ever test the theory? Patterson [1982]. It has been suggested (by Patterson 1982, among others) that the theory, or science, or methodology of classification has two main purposes:

- to express and/or reflect derivative or hierarchical relationships;
- and to provide taxonomic structures that summarize and organize knowledge.

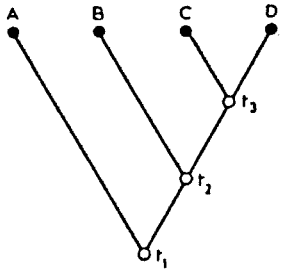
In the natural or life sciences they would probably substitute "evolutionary" for "hierarchical" above, particularly since there are extant theories such as Darwinism that have a vested interest in a particular interpretation of the derivation and origin of species. In this sense and context, classification models and theories come with their own baggage, so to speak, and particularly with Linnaean and Darwinian models that has made it difficult, if not impossible, to separate the classification methods from the value-laden theories that motivated them.

As the prefatory quote intimates, that is partially why biologists, entomologist, paleontologists and other natural scientists were so ready to seize upon the pioneering work of Hennig (1966) in a methodology of classification called "cladistics" (from the Greek word for "branching" which Hennig used to depict branching diagrams to define species and taxon relationships--see Fig. 1). The new science of cladistics offered the promise of a "value neutral" heuristic, which offered powerful and more empirical classification techniques without any of the theoretical hang-ups.[1]

Hennig himself did not necessarily want to ignore ancestral or evolutionary origins of phylogenetic relationships. In the published English version of his work, called *Phylogenetic Systematics* (1966), he defined inter-species relationships via "cladograms" or branching diagrams (see Fig. 1) that were essentially evolutionary trees, intended to capture both synchronic and diachronic data--i.e., not just to identify and summarize similarities and differences between taxa (the "knowledge representation" part of the two-part distinction that introduced this discussion), but also to reflect the common ancestries underlying the homologies so identified.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549



Hennig's definitions. Two species (C and D, for example) are more closely related to each other than to a third (A or B) if they share an ancestral species (at t_3) not shared with the third

Figure 1. A Hennig cladogram (from Patterson, 1982).

One of Hennig's many contributions was, however, to set a limit on the allowable distance that could qualify as shared ancestry. Relatively recent common ancestry ("shared derived characters") of homologous groups was more permissible than more remote ancestry ("shared primitive characters")--particularly where the former have a greater likelihood of empirical proof and support. While it is generally agreed that Hennig's Phylogenetic Systematics was set in an evolutionary framework, it can also be acknowledged that Hennig's work opened the door to questioning the established order in the theory and science of classification of the history of life.

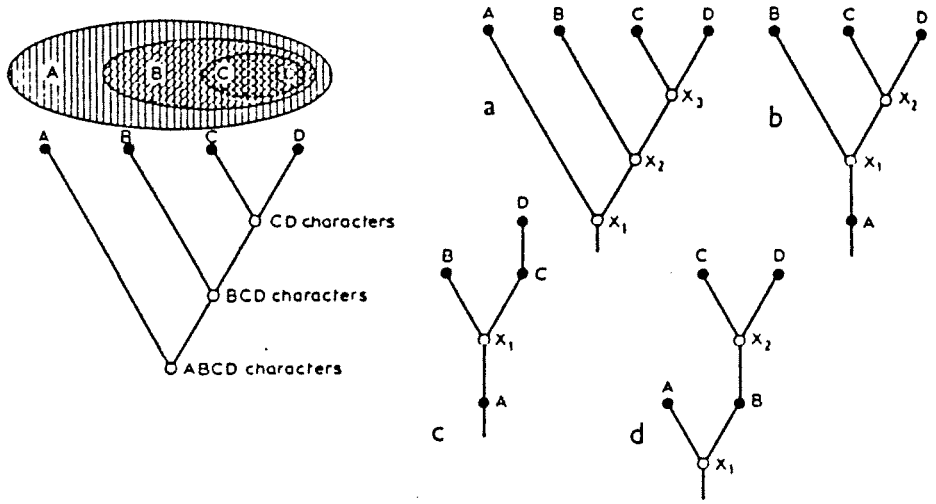
In the '70s and '80s, therefore, empiricists were responsible for more recent developments in cladistics, known as *transformed cladistics*, or *neo-cladistics* (Patterson 1982, Humphries &

1. From the perspective of a cognitive scientist and epistemologist, it seems to me that part of the reason for the controversial history of classification theories and methodologies in the natural sciences in particular, is that they have had to serve the double purpose of possessing what Chomsky would call "descriptive and explanatory power". That is, describing phylogenetic relationships is one thing (and an exercise which I believe has much in common with the task of knowledge acquisition and representation in modern expert systems), but to explain the ancestry, origin, or purpose of those relationships is quite another. Because the latter exercise is much more value-laden and law-driven it is also sometimes identified by some authors as the "nomothetic" part of classification systems, to be contrasted with the more descriptive "idiographic" task of recognizing and identifying members of a class in the first place.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

Parenti 1986), where it is possible to look at Hennig's diagrams (see Figs. 1, 2) and their associated definitions in a more general framework which has no evolutionary implications.[1]



Cladograms and trees. A cladogram is a pattern of relationships, and is exactly equivalent to a Venn diagram of sets and subsets (left). Trees (right) carry an added implication of evolution and time. The four trees are simply representatives of the 12 possible trees one could construct from the cladogram

Figure 2. Classification cladograms and trees (from Patterson, 1982).

So why is this story important, and what is its relevance to linguistics and the problem of constructing taxonomies and thesauri, or to the problem of capturing and representing knowledge in a computer?

1. Perhaps one could speak of these recent versions as 'reformed cladistics' minus the teleology of traditional cladism?--- see Ridley (1986) for a good discussion of this.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

Well, for one thing, as Patterson proposes, cladistics as a method of biological classification is a "science of pattern": "A cladogram is a summary of pattern, the pattern of character distribution, or of hierarchy in nature-- what pre-Darwinians called the 'natural hierarchy'" (Patterson 1982).

In the rest of this paper I will explore this question in the contexts of theoretical linguistics and the classification problems encountered in constructing off-line taxonomies and thesauri as well as on-line knowledge representation systems, particularly with regard to this notion of "naturalness" in classification hierarchies or derivations.

First of all, the notion of pattern seems to be crucial to all these endeavors.[1] For all of the enterprises that I have just identified, the notion of pattern is central to grouping of words/terms/descriptors/concepts on the basis of similar or differing characteristics. Hennig's cladistics have a contribution to make here for he identified 3 kinds of morphologies to describe the distribution of characters relative to a given classification/relationship problem: autopomorphies for characters unique to a species; synapomorphies for shared derived characters between groups; and symplesiomorphies for shared primitive characters within groups. For example, in Fig. 1, species C and D are "synapomorphic" because they share characters inherited from a recent common ancestor at t3.

Could we, for instance, borrow these "-morphs" of cladistics to help us bound relationships, to calculate the nearness or distance of relatedness -- all issues that come to bear on where we place a term or concept or object-definition in a node in a tree in a hierarchical classification scheme?[2] To quote Patterson (1982):" Every homology characterizes a group at some level in the hierarchy, and symplesiomorphy and synapomorphy are terms for homologies that stand in hierarchic relation: a symplesiomorphy (general character) makes a group, and a synapomorphy (special character) makes a subgroup." So Hennig provides us with a classification heuristic here when he divides the concept of similarity or resemblance into three groupings or degrees or quality of relatedness (autapomorphy, synapomorphy, symplesiomorphy), which describe the status of any character in relation to a particular problem. Surely such heuristics might address problems of class membership, or what to include or leave out in defining subject matter or building vocabularies for new domains or technologies?[3]

Likewise, the cladistic notion of "pattern" helps us take a look at the issues of degree and quality of relatedness of concepts (and even metaconcepts) from a fresh perspective, and for that alone it would be worthwhile. But the impact of traditional or neo-cladistics goes even further and deeper, because its principles of shared characters--"derived vs. primitive"-- forces us to confront

---

1. I need to make a passing comment on the phenomenon of one classification system (language) being used to construct and define applied classifications for a variety of tasks which all require the denotative (at least) use of language itself. Since language itself represents a hierarchical, multilayered system for the representation of meaning, at some point in this discussion we have to confront the metalinguistic issue of how to maintain the distinction between the signifier and the signified in classification schemata-- that is, the use of language to construct and structure systematic taxonomies whose content is itself linguistic.

2. It also comes to mind that such notions of relationship "distance" are not unlike the calculation of vector space maths that are used for pattern matching techniques in certain information retrieval systems.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

the problem that it is very difficult to construct pure, value neutral classification schemes that are not rooted or grounded in some theoretical matrix which takes a position *vis a vis* the etiology or teleology of the relationships we have identified, be those -ologies evolutionary or ancestral, or just epistemological.

For me, the notion of shared characters still raises the notion of order and precedence in systematic or structured systems. Flat, one-dimensional classification systems may vary only along a spatial continuum, but once we add levels and layers, then the dimensions of time and derivational process come into play. Whether the classification criteria are structural or functional or whatever, more complex schemata for describing, comparing, or grouping of informanda immediately start to raise the question of antecedents, of which comes first/earlier/higher; or later/lower, in our systems. For Ridley (1986) this might be termed a "pattern-process" distinction -- that is, classification *via* the pattern of distribution of taxonomic characters among species in nature, *vs.* the causal process or mechanism which "caused" the observed distribution of characters.[1]

For me, this issue seems to lie at the heart of ordered classification systems -- whether they be taxonomies (ordered on the basis of subject matter), or thesauri (ordered by categories of knowledge-- particularly Faceted Thesaurus Structures -- see the paper by Rockmore in this conference), or knowledge hierarchies in rule-based systems (ordered via semantic nets, object-based definitional hierarchies, etc.). Ultimately, the question of order and ordination -- superordination, subordination of groupings; concept class inclusion and exclusion, etc. --- means that the debate is essentially an epistemological one concerning the structure and organization of knowledge (as discussed in the second half of this paper).

I am proposing that the ongoing controversy between cladists and evolutionary systematists can be informative and instructive for our endeavors here, particularly since many of the issues they raise are generic for classification research. As Patterson (1982) points out, evolutionary systematists may agree that cladistic analysis is the best way of approaching systematics, but they claim that it does not take you far enough! Evolutionists regard production of classifications as a two-step

---

3. In my own work I am attempting to apply these notions of qualitative distinctions in taxonomic groupings to the problem of term definition and vocabulary building for technical thesauri that have to contend with newly spawned computer product terminologies. I am struck by the contrast between this work and the task of the evolutionist who bemoans the gaps in his field of knowledge and the missing data (fossil or otherwise) which bedevil his attempts to prove and justify his theories. Our task is quite the opposite in the field of computer information cataloguing. We are drowning in a sea of terminology that keeps spewing out of the high tech pipeline much like the oil from the damaged wells in Kuwait. Our task is not that of reconstruction of missing data (like the evolutionist's), but rather that of structuration of informanda in all too plentiful supply thanks to the information explosion of the second half of the twentieth century. In order to "cap" this well of information, we need systematic methods, tools, heuristics in information science, and above all a powerful, sustainable and generalizable (i.e., domain-independent) rationale and framework of information classification.

1.The kind of illustrative examples used here involving class, order, genera, species, etc., stem from the fact that most of the debate around cladistics occurs in the literature of the natural sciences. It is hoped that the reader will not have too much difficulty extrapolating or generalizing these to his or her own domain of expertise.

1.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

process--cladistic analysis followed by analysis of divergence. That is, in addition to the branching diagrams recognized by cladists, evolutionary systematists also want their classifications to take into account rate or degree of evolutionary divergence.

I suppose at this point in the discussion we could do a classification exercise ourselves, to take stock of the alternative classification heuristics that seem to be on offer here. Along a simplicity/ complexity dimension, there seem to be at least three alternatives. The simplest, purest (i.e., least nomothetic) and most parsimonious would be "phenetic classification-- where the phenotype of an organism is simply its observed characteristics, and classification of species is simply done solely according to their similarity of phenotypes. This classification meets the requirement we described earlier of descriptive adequacy, without carrying any interpretive or explanatory overhead.

The next alternative is the more traditional cladistic one preferred by Hennig -- that is, the principle of "phylogenetic classification" based principally on the phylogeny or ancestry of species-- species are grouped with those other species with whom they share their most recent common ancestor. Finally, there is the least neutral and most complicated alternative, which is the position of the evolutionists. They take all of the above but add the demand that classifications be grounded in evolutionary theory and purpose.

Part of this debate revolves around what counts as admissible evidence -- what kinds of groupings should be recognized in classifications?

For Hennig, "monophyletic" groupings are paramount -- species groupings which contain *all and only* the descendants of a common ancestor. For example, in Fig. 1, species BCD might be said to form a monophyletic group because they have a common ancestor at t2 not shared by any other taxon (e.g., A). "Such groups are called monophyletic groups and the task of phylogenetic groups is to find them."(Humphries & Parenti, 1986) This definition is more restrictive than what evolutionists can live with, but it is the preferred definitional grouping of relation types by cladists, both on empirical and qualitative grounds. The question of admissible evidence is one of defining the criteria for group identification (I am tempted to say group "membership" here, but that raises the spectre of the chicken-egg problem, where identification is a prerequisite for membership!).

Patterson (1982) points out that the rules for admissible evidence differ sharply between evolutionists and cladists: "Cladists demand that groups be characterized by synapomorphies (or homologies), so that they are monophyletic (or natural)" (1982, p. 305).

I now want to focus in on this use of the term "natural," with its assertion that one principle of grouping or classification is more natural than another. For it is this notion of naturalness that provides the bridge between cladistics and theoretical linguistics that is the real subject of this paper and to which I now wish to turn. For what we are really engaged in here is an interdisciplinary exercise about the semantics and methodologies of classification systematics (to borrow a Hennigian phrase). The first half of this paper has attempted to provide a flavor of the kind of debate that has been raging over the last 20 years within disciplines specific to the natural sciences, concerning both the theory and methodology of classification schemas. The debate is nontrivial for, as Patterson points out in our prefatory quotation, it deals with the theory and reconstruction of the history of life.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

I would like us to be able to extrapolate from this 20-years' worth of debate some methodological principles that would be theory-neutral and which we could all bring to bear on our own particular disciplines. I think that worthy goal is realizable to a large extent but, as I have pointed out above, there are always counter-claims that, phenomenological similarities notwithstanding, taxa cannot be placed and grouped in a classification system in complete ignorance of derivational characteristics. Despite Patterson's (1982) arguments, I still feel that there is some validity to the claim that all classification systematics, to the extent that they employ hierarchically ordered structures, have implicit within them, or are motivated by a theory of priority, a theory of origins. Such theories do not have to be the emotion-laden, teleological ones of evolutionism, they can be grounded in logic and epistemology, or linguistics. But nonetheless, they will bring with them presuppositions that will predetermine the definition, labelling, place and status of an entry in a classification schema. This is not necessarily a bad thing. A theory of priority can add congruence, comprehensibility and predictivity to the content and structure of a taxonomic system. The alternative is a trial-and-error approach to grouping and classification of information, which could be costly and time consuming depending on the size of the database and the purpose of the application.

I am sure one of the purposes of this conference is to be more systematic about classification schema rather than less--to find more systematic methods is to find more generalizable ones, so that the ultimate result is to have more utilitarian, multipurpose systems. And I have to confess that my bias in my own work is to be more, rather than less theory-driven.

Which brings me back to my earlier point of transition about the connectivity between certain cladistic principles and the notion of "natural forms" in theoretical linguistics. For during the 20 years that the cladistics reformation was taking place, a parallel *aggiornamento* was occurring in the treatment and discussion of a theory of "markedness" in linguistics. This topic is intertwined with a longer-lasting discussion of linguistic universals that has been ongoing in the field of linguistics and language development for several decades. It has been treated by Jakobson (1941), and Greenberg (1966), and touched upon by McCaffrey (1971), and by Chomsky and Halle (1968) in their textbook *Sound Pattern of English*. But the subject seemed to lie dormant during the '70s, followed by a re-awakening of interest and research in the 1980s (e.g., Battistella 1990; Andrews 1990), and I now want to draw upon that recent body of literature in order to make the kind of inter-disciplinary synthesis of knowledge and information that I am trying to grapple with in this paper.

The kind of parallels I want to draw are as follows: Markedness has to do with the naturalness of linguistic forms. Naturalness, in turn, has to do with primacy and order and status derivation and with simplicity/complexity and generality/specificity distinctions concerning contrastive linguistic forms, viewed from both a historical and a usage perspective. Whether these polar oppositions are phonological or syntactic or semantic, the same arguments apply. Whether in occurrence or usage (synchronic linguistics) or in historical evolution and development (diachronic linguistics), one partner/half/version of a polar opposition may be characterized as "unmarked," more natural, more general, more widely distributed, etc., while the other half will be defined as "marked", less natural, more specific, less widely occurring, etc. With the help of Battistella (1990) and my own research, I have constructed Table 1 to illustrate what some of these markedness distinctions are.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

Table 1.
CHARACTERISTICS OF UNMARKED vs. MARKED FORMS IN LANGUAGE

UNMARKED	MARKED
more natural	less natural
more general, simpler	more specific, complex, focussed
more widely distributed in languages of world	less widely distributed
more frequent	less frequent
learned earlier/first by children learning a language	learned later (if it is a meaningful contrast in their language
less "cost" (-value)	more cost (+value)
"ground"	"figure"
less context-dependent	more context-dependent

Battistella (1990) provides some good examples of grammatical and semantic markedness in everyday language usage, and there are phonemic examples also. Examples deal with contrasts between singular vs. plural, past vs. present tense, lexical oppositions about physical qualities (age, height), etc. For example, consider the lexical contrasts in the following statements:

1. a) How old are you?
   b) How young are you?

2. a) How tall are you?
   b) How short are you?

As Battistella (1990) points out, the (a) examples imply nothing specific about the age or height of the addressee, while the (b) examples do. In these kinds of lexical oppositions we can see the notion of markedness playing two roles, or "marking" two kinds of oppositions. In the above pairs of words referring to opposite physical qualities, the unmarked member of a word pair (old, tall) can play the role of either referring to the all encompassing and subsuming general concept, or it can represent one particular value of the general property. In Battistella's examples the unmarked concepts "old" and "tall" in (a) refer to general properties *age* and *height* with no implication about specifics; OR they can be used to refer to specific contrastive values opposite to *young* and *short*.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

So the more natural, unmarked element can signal a presence/absence of a contrast, or denote a specific contrast. This illustrates the two major roles that the concept of markedness plays in linguistic analysis. First, it plays a pragmatic, functional role as a refiner of meaning in usage--it helps flag distinctions in meaning. Second, it provides an analytic tool as an evaluative metric with its binary scoring of (-)/Unmarked and (+)/Marked values for linguistic contrasts. These functions that markedness provides help constitute, in my estimation, a "virtual etiology" for classifying linguistic features on the basis of both usage and primacy. In this respect they are comparable to cladistic classification requirements which need not only to construct taxonomies which capture and organize available knowledge, but also to place it in a synchronic and diachronic context.

So one point of connection that I want to make here is a structuralist one, *viz.* that the concept of naturalness or markedness serves as an organizing principle for language patterns. I started out my discussion of cladistics by suggesting that at least one of the purposes of a classification system was to structure and organize knowledge. There is a similar structuralist origin and underpinning to the recent treatment of markedness by Battistella, both historically and theoretically. For example: "The principle of structure expressed by the concept of markedness is that, whenever we have an opposition between two things, one of those things--the unmarked one--will be more broadly defined" (1990, p.4).

The one significant difference between the kinds of structural constraints encountered in man-made or intentional classification schemes and the structural dimensions dealt with in the treatment of linguistic markedness is that the latter phenomena are more "organic" or intrinsic to the nature of language. We are here entering into the territory of semiotics and the organization of sign systems, which, in certain theoretical formulations, might encompass language-specific usage patterns, or language universal patterns, or both. What is at issue is the functionality of markedness for language. For reasons of internal consistency and coherence, for reasons of the systematicity inherent in language, as well as for historic-linguistic reasons, the case can be made that "language exhibits congruence between the markedness of meanings (signifieds) and the markedness of expressions (signifiers)"(Battistella 1990).

Consider some further examples of the relation between expressions and underlying meaning. In addition to the kinds of lexical/semantic oppositions exemplified above, Battistella also cites grammatical examples of opposition between past and present tenses which provide a good illustration of the two critical aspects of markedness notation: function and value.

3. a) I am stepping through the door (happening now)

   b) I wear sneakers (habitual)

   c) I leave for the continent next week (near future)

   d) So then I say to him: "Shut up!" (past)

In these examples, Battistella claims that present tense is unmarked and unspecified with respect to time, covering past, present, future; while the past tense is marked with respect to present.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

I can summarize these distinctions in the following diagram:

DIMENSION:	TIME	
OPPOSITIONAL CHARACTER/CONTENT:	PRESENT	PAST
VALUE:	UNMARKED(-)	MARKED(+)

"The past tense marks past time; the present is unmarked, and its actual time reference depends on context or on other semantic properties of the verbs in question." Further, "the present tense....is the general, enveloping, subsuming tense, as opposed to the more restrictive and focussed past tense" (Battistella, 1990, p.3).[1]

Similar evidence can be provided for phonological contrasts involved in the speech sound distinctions of a language which establishes their fundamental property as vehicles of meaning. Table 2 shows how the phonemes of a given language (the columns) can be valued along a binary, polar dimension, captured as a set of distinctive features or attributes (the rows of the table) which are more universal in nature. Table 3 then shows the markedness status of the same phonemic contrasts. So we first have the traditional linguistic classification of phonemes according to distinctive feature analysis (Table 2), then we map onto those same phonemic segments a value for the naturalness (or not) or basic-ness (or not), scored as U or M, of the particular contrast (Table 3).

Table 2

Distinctive Feature Classification

	p	t	č	k	b	d	ǰ	g	f	θ	s	š	h	v	ð	z	ž	m	n	ŋ	l	r
consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
vocalic	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+
nasal	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−	−
compact	−	−	+	+	−	−	+	+	−	−	−	+	+	−	−	−	+	−	−	+	−	+
abrupt	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	+	+	+	+	+
strident	−	−	+	−	−	−	+	−	−	−	+	+	−	−	−	+	+	−	−	−	−	−
tense	+	+	+	+	−	−	−	−	+	+	+	+	+	−	−	−	−	−	−	−	−	−
grave	+	−	−	+	+	−	−	+	+	−	−	−	−	+	−	−	−	+	−	+	−	−

1. Battistella goes on to suggest that the more natural, unmarked term of an opposition serves as a "conceptual default value" that is assumed, unless the context specifies otherwise--then carries this further with the intriguing notion that, in the gestalt of a total utterance, the unmarked term serves as the "ground" against which the marked term appears as "figure" (1990, p.4).

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

Table 3

Markedness Values of Consonants

	p	t	č	k	b	d	ǰ	g	f	θ	s	š	h	v	ð	z	ž	m	n	ŋ	l	r
consonantal	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
vocalic	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	M	M
nasal	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	M	M	M	U	U
compact	U	U	M	M	U	U	M	M	U	U	U	M	M	U	U	U	M	U	U	M	U	M
abrupt	U	U	U	U	U	U	U	U	M	M	M	M	M	M	M	M	M	U	U	U	U	U
strident	U	U	M	U	U	U	M	U	M	M	U	U	M	M	M	U	U	U	U	U	U	U
tense	U	U	U	U	M	M	M	M	U	U	U	U	U	M	M	M	M	U	U	U	U	U
grave	M	U	U	M	M	U	U	M	M	U	U	U	M	M	U	U	U	M	U	M	U	U
composite value	1	0	2	2	2	1	3	3	3	2	1	2	4	4	3	2	3	2	1	3	1	2

(Tables 2 and 3 from Battistella, 1990.)

The distinctive feature matrix by itself is akin to the taxonomic representations of transformed cladistics, with its attempt to be theory-neutral in its biological comparisons. When you place beside it (or perhaps even orthogonal to it?) the markedness matrix (Table 3), I would claim that you have added an evaluative dimension and that you can no longer avoid the determining issues of primacy, order, and the "naturalness" of derived forms--- every linguistic opposition, be it phonological or syntactic or semantic, forces the same question about linguistic antecedents: in the evolution of linguistic forms, which form/feature came first, earlier, etc.? (and if form denotes meaning, and meaning denotes concepts, then what we are really talking about is a descriptive framework for the structural representation of inheritance and subsumption in knowledge schema.)

Interestingly enough, Battistella subtitled his book on markedness, *Evaluative Superstructure of Language*, and it is this central notion of evaluation that brings this discussion of linguistics closer to some of the debate in the cladistics literature. Our use of language in speech presupposes an evaluation of our choice of terms -- evaluation of relevance, of appropriateness, of emphasis, of continuity, etc. Likewise, our use of language as a tool in classification systems also presupposes an evaluation of the use of a term or concept for its relevance, power, generality, subsumption, etc. What the theory of markedness suggests is that such usage is not arbitrary or ad hoc, but, on the contrary, is motivated by principles. It is one of the claims of this paper that uncovering and applying such principles would help create a classification systematics that could be applied to many domains.

To summarize some of the linguistic points so far: Linguistic units of analysis -- be they phonological, lexical, or grammatical-- can be construed as being constructed of groupings of features whose constellation in particular utterances represent and convey basic meaning

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

distinctions. Examples are plentiful. In Phonology, for example, as can be seen in Tables 2 and 3 above, one kind of opposition for consonants can be in the feature "nasal/non-nasal" (remember Woody Allen in "Take the Money and Run" scribbling his robbery note to the bank teller as "I have a gub" !). Lexical opposition was discussed in the age/height examples earlier. Similarly, grammatical subdivisions and oppositions of features would include the contrast of nouns vs. pronouns-- then the further subdivision of these into singular vs. plural--then further subdivision of pronouns into nominative vs. objective -- and so on.

The point to be made here is that the theory of markedness builds on these notions of distinctive features and oppositions which are central to the study of both linguistics *and* systematics.

So these are critical issues in the contemporary discussion of the role of markedness in linguistic analysis (Andrews 1990), which provide connectivity with similar debate over classification methodology in cladistics, with its concern over homologies, groupings, relationships, ancestry or derivations, etc. In the linguistics of markedness, the counterpart concepts are oppositions, features, patterning, systematic relationships and hierarchical structure.

Let me conclude this discussion of the linguistics-cladistics connection by dwelling for a moment on these latter notions of patterning and hierarchy in a theory of language, for I think this will help pull some of the interdisciplinary threads together in this argument that I am weaving.[1]

First, the structuralist viewpoint that language must be studied in its totality, taking into account the symmetry, balance and inter-relatedness of both its structure and function. We are here dealing at a metalevel, for language is a metaphor for one of the most structured systems known to man. So it is an appropriate model for the study of systematics, which is germane to many disciplines, including cladistics and classification research.

A structuralist analysis of language-as-system reveals that it contains forms (signifiers) and meanings (signified), and that it uses several layers of structures, organized hierarchically but functioning in real time as parallel multiprocessors, to support its purpose as a communicator of meaning. This dual role is carried by the primary structural divisions within language, usually recognized as phonology, syntax and semantics.

> Phonology and semantics comprise a duality of patterning in the sense that language is a system of material signifiers linked to a separate system of conceptual signifieds, a system of forms linked to a system of meanings. The signifiers are units of linguistic form -- distinctive features, phonemes, morphemes, words, phrases, and utterances. The signifieds are the corresponding units of meaning. Each of these systems is defined by a set of features. Features thus create networks of signifiers or

1. For the kinds of viewpoints espoused here, Battistella (1990) and I both draw heavily on the Structuralist school of linguistics represented by the Prague School and Roman Jakobson. For a worthwhile treatment of the history of this school, see either Battistella (1990) or Andrews (1990).

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

> signifieds by encoding oppositions. The focus of study on the plane of signifiers is with features of linguistic expression, while the focus of study at the level of signifieds is on features of meaning (Battistella 1990, p.19).

I would add a further clarification here that the "network of signifieds" is a conceptual effect, not a cause, of the operation of encoding feature oppositions in utterances.

This powerful notion of networks of forms and meanings co-exists with the notion of language as a complete hierarchical system -- a system where parts and wholes co-habit, and where qualitative distinctions within and between them define relationships (and ultimately, meanings). Features partake of this hierarchiization too. As I tried to argue in my discussion of Table 3, features can be ranked in a hierarchy of implication and evaluation, where certain features (more natural, more primary ones) are superordinate to other subordinate features or are implied by them.

Graphically one could represent a feature hierarchy via a distinctive feature matrix (Table 2), or use a tree diagram to reflect oppositions ranked *via* superordinate vs. subordinate position. The point can also be made that there are sufficient practical examples of the hierarchical nature of linguistic categories that are at least implicit in everyday manifestations:

> The hierarchical nature of lexical and grammatical categories is well established. In lexical analysis, hierarchy is revealed in part through taxonomies of concepts (as in the heuristic taxonomy implied in the organization of a thesaurus or in the formalized meaning postulates or redundancy rules of linguistic semantics). And in traditional grammar, hierarchy is implied by the fact that words and grammatical constructions are listed in dictionaries and handbooks under a main or citation form: one form is taken to be basic, and the others are treated as derived forms (Battistella, 1990, p.20).

So we have concepts and meanings capable of being structured hierarchically. And we have layers of linguistic forms that can be similarly organized. Then we have language itself as the supersystem, the prime model for a science of systematics which can be used both by those who classify (e.g the cladists) and those who design and build classification schema.

And we must not forget the role of hierarchic structure in defining "relationships," the concept so important to the science of pattern in cladistics. Hierarchical systems based on linguistic principles not only place terms in a superordinate or subordinate position, but they also carry within them a dynamic that drives the definition and determination of relations and their properties.

Let Battistella have the last word on the inter-relationships of hierarchies, oppositions, and markedness in language:

> Hierarchy is reflected in the dominance of more general terms over less general, and we can view markedness as a hierarchization of opposites. The concepts of markedness, opposition, and hierarchy are thus intrinsically linked. Opposition imposes a symmetry or equivalence upon language: within a minimal paradigm two signs are defined by the presence versus the absence of a property. Hierarchy is an

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

> evaluative component that organizes related categories. Markedness is the projection of hierarchy onto the equivalence implied by opposition, extending the non-equivalence principle of a ranked taxonomy to the minimal oppositions that make up the quanta of language (Battistella, 1990. p. 20).

So markedness pervades the notion of "value" throughout the hierarchical system -- the very notion that was so controversial in the cladistics *vs.* evolutionists debate. But in the case of language it does so in a way that seems to me more soundly based on principle, on empirical evidence of usage, and on more cohesive theoretical foundations.

So can cladistics learn from Linguistics?--and can we at this conference learn from both? Let me conclude by offering a couple of points of connectivity from my own perspective. For the range of topics and interests that are represented at this conference, I think that cladistics and Theoretical Linguistics both offer models, heuristics, concepts, definitions and tools for anyone designing classification schema. I have already suggested that recent developments in these fields can provide the foundation for a science of "Classification Systematics". Next, for those of us who have to build vocabularies and thesauri for exotic terminologies, I think that structural linguistics, and the theory of markedness in particular, provide structures and rules to help guide the difficult task of using linguistic forms to classify linguistic content. I see a particular connection between the notions of opposition, hierarchy and markedness values and the task of categorizing superordinate concepts and classes of knowledge that is encountered in the construction of faceted thesaurus structures (see, e.g., Rockmore 1991, and also her paper at this conference).

Further, I also see connections with current work on "Super Thesauri" and metaknowledge organization (Stern & Rischette 1991). Stern's work employs concepts of object-oriented design to deal with problems of hierarchical and equivalence relationships in classes, superclasses and metaclasses. There seems to me to be obvious connections between such research and the topics treated in this paper, particularly regarding the hierarchically structured system models and relationship criteria that both cladistics and Linguistics provide.

Finally, in my own work on knowledge engineering, I know from experience that the young specialties (they are not yet sciences!) of Artificial Intelligence and Expert Systems offer many toolkits and approaches to the complex task of capturing and representing knowledge in a computer, but no theories which meet Chomsky's criterion of minimal descriptive adequacy, let alone explanatory power. These recent approaches have had limited success with certain narrowly defined fields of expertise, but have made no progress in dealing with the problem of how to represent linguistic knowledge in a computer. They offer a catalogue of techniques for knowledge representation-- scripts, frames or slot-and-filler methods, semantic nets, object-oriented definitional paradigms, rule-writing paradigms, and so on--- but there is no unifying frame of reference or theory that would motivate a program of systematic research on this problem.[1]

---

1. Lest my critique here sound too pessimistic, I won't deny that some of these new methodologies from the field of Artificial Intelligence are useful or relevant. For instance, I can see parallels between the operation of backwards chaining in expert reasoning systems and the methods the cladists use to formally trace, depict, and account for ancestral derivation of taxa. The point of my critique is that such A.I. techniques represent a technology, not a science or a theory.

ISSN: 2324-9773

McCaffrey, A. (1991). Applied Cladistics: New Models for Classification and Taxonomy Research;or How the New York Review of Books Taught Me Everything I Needed to Know about Taxonomy Research. 2nd ASIS SIG/CR Classification Research Workshop, 81-96. doi:10.7152/acro.v2i1.12549

But perhaps we should not even expect technology-driven disciplines to solve their own problems. I feel more confident that the inter-disciplinary synthesis of theories and methodologies that I have been pursuing in this paper will provide a more fruitful path to the solution of the classical problems of knowledge acquisition, organization and representation that underlie most of our interests at this conference.

BIBLIOGRAPHY

Andrews, E. 1990. *Markedness Theory.* Durham, N.C.: Duke University Press.

Battistella, E.L. 1990. *Markedness: The Evaluative Superstructure of Language*. Albany, NY: SUNY Press.

Chomsky, N., Halle, M. 1968. *Sound Pattern of English.* NY: Harper & Row.

Greenberg, J. 1966. *Language Universals.* The Hague: Mouton.

Hennig, W. 1966. *Phylogenetic Systematics.* Urbana, IL: University of Illinois Press.

Humphries, C.J, Parenti, L.R. 1986 *Cladistic Biogeography.* Oxford: Clarendon Press.

Jakobson, R. 1941 (1968). *Child Language, Aphasia and Phonological Universals.* The Hague: Mouton.

McCaffrey, A. 1971. *Speech Perception in Infancy.* Doctoral Dissertation. Cornell University, Ithaca, NY.

Patterson, C. 1982. Cladistics and classification. *New Scientist,* 94: 303-306.

Platnick, N.I. 1979. Philosophy and the transformation of cladistics. *Systematic Zoology,* 28: 537-546.

Ridley, M. 1986. *Evolution and Classification: The Reformation of Cladism.* London: Longman.

Rockmore, M. 1991 Facet analysis and thesauri for corporate information retrieval. In: *Advances in Knowledge Organization. Vol. 2.* Frankfurt: Indeks Verlag.

Sokal, R.R., Sneath, P.H. 1963. *The Principles of Numerical Taxonomy.* San Francisco, CA: W.H. Freeman.

Stern, A., Rischette, N. 1991. On the construction of a super thesaurus based on existing thesauri. In: *Advances in Knowledge Organisation. Vol. 2.* Frankfurt: Indeks Verlag.

ISSN: 2324-9773