Taxonomy Revision in Botany: A Simulation of Historical Data *

Eugenio Alberdi and Derek Sleeman

Computing Science Department
University of Aberdeen, Scotland
{eugenio, sleeman}@csd.abdn.ac.uk

1 Introduction

In this abstract we present ReTAX, a system for *taxonomy revision* in Botany. The function of ReTAX is to revise, and eventually modify, a taxonomic hierarchy as novel inconsistent data are provided. The system has been applied to replicate some taxonomic revisions which have taken place historically in the botanical family Ericaceae.

As noted by different researchers (e.g., Lakatos, 1976; Kulkarni & Simon, 1988; Klahr, Dunbar, & Fay, 1990), focusing on *unexpected observations* is a powerful heuristic in scientific reasoning. Since classification is the basis of *many* scientific endeavours (see, e.g., Sokal 1974), studying the effect of unexpected phenomena on the revision of classificatory systems is of particular relevance for scientific discovery. Some artificial intelligence models of discovery have dealt with the creation of scientific taxonomies (e.g., Langley, Zytkow, Bradshaw & Simon, 1983; Nordhausen & Langley, 1990). Likewise some computational systems have been created to perform taxonomy related tasks such as biological identification, automation of taxonomic descriptions, database management, and various numeric analyses (see, e.g., Pankhurst, 1975; Abbot, Bisby, and Rogers, 1985; Woolley & Stone, 1987; Watson, Dallwitz, Gibbs, and Pankhurst, 1988; Domingo, 1993). However, in general, neither of these approaches has addressed explicitly taxonomy revision. On the other hand, discovery systems which have tackled the revision of faulty scientific theories to accommodate anomalous data (e.g, Rose, 1989; Rajamoney, 1990; Karp, 1993) have not normally dealt with classification. ReTAX applies a theory revision approach in the context of scientific classification.

ReTAX is based on a psychological study on *category induction*, in the domain of plant taxonomy, that we conducted at Aberdeen. The purpose of the study was to detect the search strategies and heuristics used by expert botanists when facing puzzling phenomena as they solved a categorisation problem. The protocols resulting from this study were modelled computationally and were used subsequently to define some of the revision mechanisms incorporated in ReTAX. Before we describe ReTAX and its application to historical botanical data (in Section 3), we give in Section 2 a brief description of the psychological study and the associated computational simulation.

* We would like to thank Dr. G. Smith for his advice and assistance with the selection of botanical data, Mrs. C. Green for her help with the rating of the protocols, Dr. R. Logie (from the Psychology Department at Aberdeen) for his support and guidance on the psychological aspects of the research, and the expert botanists which agreed to take part as subjects in the psychological study. Mr. E. Alberdi was supported by a research studentship awarded by the Spanish Government (Ministerio de Educación y Ciencia).

2 Psychological Study and Simulation in the Domain of Plant Taxonomy

The task in our psychological study can be viewed as a *"laboratory" simplification* of some of the processes involved in taxonomy revision. In particular, we were interested in those situations where a botanist searches for new discriminant descriptors to accommodate a new item which is inconsistent with an established taxonomy.

We based our psychological work on research carried out by Korpi (1988) on inductive categorisation. Korpi investigated the effects of presenting unexpected items on a category identification task which involved everyday categories and natural concepts. We adapted Korpi's empirical approach to the constraints of a real scientific domain, with scientific data and expert scientists.

In our study of category induction, five expert botanists were faced with the task of identifying a botanical category from a set of positive and negative examples. Each example consisted of a graphic illustration of a species.

In order to create a puzzling situation in this task, we followed essentially the same strategy used by Korpi (1988) in her category identification study. We selected the stimuli according to two kinds of features: a dominant feature and a subsidiary feature (as illustrated in the example shown in Figure 1). A "dominant feature" ("siliqua type of fruit" in the example) is a characteristic which is discriminant, obvious, and easily observable, so it is a feature which a botanist is expected to pay primary attention to when observing a plant. A "subsidiary feature" ("entire leaves" in Figure 1) is a characteristic which is not so obvious as a dominant feature; it is *not* normally used, in botanical classification, as the major characteristic that describes a taxon. In our study, the feature that determined the membership of an instance to a given category was the "subsidiary feature", while the "dominant feature" was

	Arabis hirsuta	*Coringia orientalis*	*Coronopus didymus*	*Erophila verna* (Rogue)	*Matthiola incana*	*Teesdalia nudicaulis*	*Cardamine pratensis*
DOMINANT FEATURE: **Fruit: siliqua**	+	+	-	-	+	-	+
SUBSIDIARY FEATURE: **Entire leaves**	+	+	-	+	+	-	-
CLASSIFICATION	+	+	-	+	+	-	-

Figure 1: *Examples and counterexamples of category "flowers with entire leaves" in our psychological study.*

used as a misleading characteristic.

The items shown to the subjects for each category were arranged in such a way that the first positive examples that the subject saw (e.g., *Arabis hirsuta* and *Coringia orientalis* in Figure 1) shared both the "dominant" and the "subsidiary" features while both features were missing in the initial negative instance(s) presented (*Coronopus didymus* in our example). It was expected that the subject would start forming her/his categories on the basis of the "dominant feature", probably ignoring the "subsidiary" one. After a few examples consistent with respect to the "dominant feature", we introduced a "rogue", which was a positive example that lacked the "dominant feature" but possessed the "subsidiary" one (*Erophila verna* in Figure 1).

The study followed the procedures typical of think-aloud protocol elicitation (Ericsson and Simon, 1984); we recorded and analysed the verbal reports elicited by the subjects as they were working through the categorisation task. The analysis of the protocols suggests a number of search strategies and heuristics which support a focusing model of categorisation consistent with the results of Korpi's (1988) study. The model is characterised by the interaction of inductive search and expert domain knowledge. The inductive procedures used by the subjects were reflected mainly in exhaustive observation and comparison of the visual items (graphical illustrations). The expert knowledge interacted with the inductive processes by helping the subjects focus on the relevant information. In particular, this interaction of the taxonomic knowledge with the data was reflected in:

a) *A search in the data for features which are relevant for taxonomic purposes.* When observing and comparing the items in the study, all the subjects tended to mention the same characteristics first. These features coincided with the descriptors highlighted in standard taxonomic practise as discriminant features.

b) *Refocusing the search after an impasse.* When an item (normally the "rogue") invalidated a thread of reasoning, the subjects started exploring new alternative features. The source of the new features was often the knowledge possessed by the subjects about the hierarchical relationships between the items and about the different degrees of discriminability of the botanical descriptors.

The effect of presenting unexpected items in our study was, therefore, a shift of focus in the search for descriptors at a hypothetical level. More details of our psychological results can be found in Alberdi & Sleeman (to appear).

The results of our psychological study were modelled in a computer programme which was given the same task as the expert botanists. The programme was given as input a taxonomic hierarchy containing information about the family Cruciferae. The type of information implemented was similar to the information given as input to ReTAX (see Section 3.1). The simulation programme used knowledge about the discriminability power (or dominance) of the descriptors in family Cruciferae to explore the botanical hierarchy and find a common link between a group of botanical items (in particular, the set of examples represented in Figure 1). The first hypotheses generated by the programme contained features which were considered "dominant" for the classification of the items and their corresponding parents in the hierarchy. When a "rogue" was presented and the initial hypotheses were invalidated, the search switched to alternative, less dominant features, until the common link between the items was identified. Using this shift of focus in the search for descriptors, the programme managed to replicate many of the behaviours elicited by the botanists during the psychological study.

3 ReTAX: A System for Taxonomy Revision

3.1 General description of ReTAX

ReTAX is designed to interact with a user as it performs the revision of a taxonomic hierarchy. The user provides the new items ReTAX has to classify and the values for the descriptors that characterise the items. The underlying assumption of ReTAX is that the botanical hierarchy that it is given initially is inconsistent with the taxonomic knowledge possessed by the user. It is assumed that one of the major sources of inconsistency between ReTAX's and the user's taxonomic knowledge is because the features that ReTAX considers relevant for the classification of the items ("dominant features") do not coincide with the features considered "dominant" by the user. A principal function of ReTAX is finding the discriminant descriptors

which permit the system to classify the items consistently with the taxonomic knowledge of the user.

ReTAX is given as *input*:

1) A frame-based botanical hierarchy with the following information:
   - "is-a" relationships between the taxa,
   - The set of values of different descriptors for each of the taxa (some of the values and descriptors are inherited from parent taxa)
   - The dominance (discriminant power) of each descriptor for each taxon, expressed in a numeric index (*discriminability index*). To be consistent with the analysis used in our psychological study, ReTAX uses the discriminability indices to recognize two different types of descriptors: dominant features and subsidiary features. A dominant feature which discriminates two particular taxa will be referred to as a *distinctive feature*.
2) A set of taxonomically relevant descriptors which are not instantiated in the hierarchy (*independent taxonomic features*).
3) A set of "consistency checking rules" (see Table 1.a).
4) A set of refinement operators (see Table 1.b).

ReTAX is presented with a new item which is classified by the user as a member of a given species. ReTAX attempts to update the given hierarchy by asking the user for values of the features that the system considers dominant. Then ReTAX checks whether the information associated with the new item is consistent with the taxonomic knowledge it has stored. ReTAX detects an inconsistency if either of the "consistency rules" (shown in Table 1.a) is violated.

If no inconsistency is detected, ReTAX simply stores the new item with the appropriate values. If an inconsistency is detected, ReTAX reports it, and asks the user if he/she wants to alter the information he/she has provided. If the user confirms the initial information, ReTAX stores the item as a conflictive item. If, after trying to classify new items, ReTAX encounters more conflictive cases, it applies the pertinent refinement operators. Table 1.b shows a list of the seven most important refinement operators implemented in ReTAX.

After applying a refinement operator, ReTAX checks whether the two "consistency rules" (in Table 1.a) are met. If there is still an inconsistency, ReTAX tries further refinement operators. If, as a consequence of applying a refinement operator, new inconsistencies are generated, ReTAX backtracks and tries alternative operators. This cycle is repeated until all the known items are consistently classified in the revised hierarchy. In some cases more than one plausible refinement will be possible. In such cases, the user is asked to choose the most appropriate refinement. If ReTAX fails to classify the new items consistently, this failure is reported to the user.

Below we describe an example of ReTAX's performance as it tackles the revision of the botanical family Ericaceae. In this example we show how ReTAX is able to replicate some of the taxonomic refinements performed historically within that family for taxa *Pernettya* and *Gaultheria* (Middleton & Wilcock, 1990).

3.2 Taxonomic history of genera *Pernettya* and *Gaultheria* in the botanical family Ericaceae

One of the earliest and most influential studies of the family Ericaceae was made by Hooker (1876). In Hooker's classificatory schema (a subset of which is reproduced in Fig. 2), *Pernettya* and *Gaultheria* appear as two separate genera. The main differences detected between the two genera were the "type of fruit" and the "succulence of the calyx" that surrounds the fruit. The fruit in *Gaultheria* is normally a capsule surrounded by a succulent calyx while in *Pernettya* the fruit is a fleshy berry with a dry (non succulent) calyx. Although Hooker was aware that some species of *Gaultheria* were very similar to some of the species of *Pernettya*, he still maintained the distinction between the two genera and even classified them as belonging to different tribes, because he gave a big emphasis to fruit characteristics. Taxonomists after Hooker, looked at a wider range of taxonomic characteristics (specially leaf anatomy) and modified this early scheme (Middleton & Wilcock, 1990). Most of these newer schemes do not group *Pernettya* and *Gaultheria* in two different tribes, but still keep them as separate genera because of the mentioned fruit differences. Some taxonomists, though, realised that these fruit characteristics did not discriminate between the genera. Some other taxonomists mentioned, as a difference between the two genera, two other features ("flower dioecism" and presence of "vivipary") which had been reported for some species of Pernettya and had never been detected in any species of Gaultheria (see Middleton & Wilcock, 1990).

Middleton and Wilcock (1990), of the Plant Science Department at Aberdeen, carried out a further study of the Ericaceae. They reported several species of *Gaultheria* and *Pernettya* which contradict the division of the two genera on the grounds of the characteristics mentioned above. In this study, Middleton and Wilcock analysed multiple (more than 100) morphological and anatomical features in 118 species of the two genera. They concluded that *Pernettya* cannot be maintained as a separate genus and reclassified all the species of *Pernettya* as members of *Gaultheria*.

a) CONSISTENCY CHECKING RULES

R1.- All the siblings at a particular level must have a unique set of instantiated[1] dominant features. Each class will be distinguished from another sibling by:

1a. having a distinct set of values for at least one of the instantiated dominant features they share

1b. being the exclusive parent of its subclasses

R2.- Each feature of a class must either be:

2a. the same as the corresponding feature in the superclass

2b. *or* a specialisation of the corresponding feature in the superclass

b) LIST OF REFINEMENT OPERATORS

Feature-updating Operators

OP.1 - Generalise Descriptor

CONDITION: When a descriptor in a given item is not the same or a specialisation of the corresponding descriptor in the class the item belongs to (violation of Consistency Rule 2).

ACTION: Enlarge the set of values for the corresponding feature in the class so that it covers the given item. IF after generalising a feature, that feature stops being discriminant:

-> Remove generalisation and store item as conflictive

OP.2 - Decrease Discriminability Index of Descriptor

CONDITION: When a given dominant feature does not discriminate between two classes.

ACTION: Decrease the discriminability index of the feature for those classes.

OP.3 - Remove Dominant Descriptor(s)

CONDITION: When the observation of one or more inconsistent items have forced ReTAX to lower the discriminability index of a descriptor (OP.2), and this index is below a given threshold (T).

ACTION: Treat the descriptor as a "subsidiary feature".

OP.4 - Search for New Dominant Descriptor

CONDITIONS:

- When an item can belong to more than one class at the same time (violation of Consistency Rule 1b).
- When all the "dominant features" of a given class have been removed (OP.3).

ACTION: Search, among the subsidiary features (and the *independent taxonomic features*), for a feature which facilitates the discrimination between the conflictive classes. Compare the values for the mentioned features among the members of the conflictive classes.

OP.5 - Increase Discriminability Index of Descriptor

CONDITION: When a feature, originally considered "subsidiary", is detected as discriminatory between the members of a class (after applying OP.4).

ACTION: Increase the discriminability index of the feature to a value above the threshold T.

Hierarchy Re-structuring Operators

OP.6 - Move Class DOWN in the Hierarchy

CONDITION: When two classes at the same level of the hierarchy *cannot* be distinguished from one another by any significant dominant feature (OP.4 has failed).

ACTION: Classify the smallest class as a subclass of the largest class.

OP.7 - Move Class UP in the Hierarchy

CONDITION: When a class cannot be successfully classified as a child of its corresponding superclass (R2 violated) or as a child of the sibling classes of its corresponding superclass (failure of OPS. 1, 4, 6).

ACTION: Change the rank of the class and place it at the same level as its immediate parent.

1. By instantiated feature we mean one which has been specialised by giving it a particular value. Also the top level class might specify the feature "legs", and the corresponding child would instantiate the value "legs: 3".

Table 1. *Consistency checking rules and refinement operators in ReTAX.*

3.3 Performance of ReTAX

In order to reproduce the taxonomic refinements described above, we implemented in ReTAX a subset of the classification scheme produced by Hooker (1876) (represented in Figure 2). We implemented in the hierarchy information about the ranks "family", "genus" (plural: "genera"), and "species", which are conventionally considered to be the most taxonomically relevant ranks in Botany (Davis & Heywood, 1964). We used a total of about 75 descriptors (with different degrees of "dominance") to characterise the different taxa in the hierarchy. As in Hooker's schema, the distinctive dominant features between genera *Pernettya* and *Gaultheria* were "type of fruit" and "calyx succulence in fruit".

ReTAX was presented with twelve new items: seven species of *Gaultheria* and five species of *Pernettya*. The first five items were three species, traditionally classsified as *Gaultheria*, which had a fleshy fruit (berry), and two species, traditionally classified as *Pernettya*, with a dry fruit (capsule). This information violated the "consistency rule" R2 (in Table 1.a); so ReTAX attempted unsuccessfully a generalisation of feature "type of fruit" in both *Gaultheria* and *Pernettya* (application of OP.1 in Table 1.b). Since the feature did not discriminate the two genera any more (violation of rule R.1.a) the discriminability index of the descriptor was reduced in both genera (OP.2) until it was eliminated as a distinctive feature (OP.3).

Pernettya and *Gaultheria* were still separated as two distinct genera because of differences in the feature "calyx succulence in fruit". But four new items were presented (two species of *Pernettya* and two species of *Gaultheria*) which contradicted the distinction of the genera on the basis of this feature. After applying the same operators as above (OP.1, OP.2, and OP.3), the feature "calyx succulence" was removed as a dominant descriptor for the genera. ReTAX then searched for new discriminatory features (OP.4) among those features considered subsidiary and among the *independent taxonomic features*. The system encountered one new dominant feature: presence of "vivipary". All the species of *Pernettya* known to the system were viviparous, while none of the *Gaultherias* was. This feature was given a new discriminability index (OP.5).

Again, three new items were presented. One of them was a species of *Gaultheria* with "vivipary". ReTAX, after a failed generalisation of the feature (OP.1), decreased its discriminability index (OP.2). Since the feature "vivipary" had just been acquired, its discriminability index was very low and the descriptor was eliminated as a dominant feature (OP.3).

A new exhaustive search among the subsidiary features and the *independent taxonomic features* (OP.4) was executed and ReTAX concluded that no new descriptors could be found to distinguish the two genera.

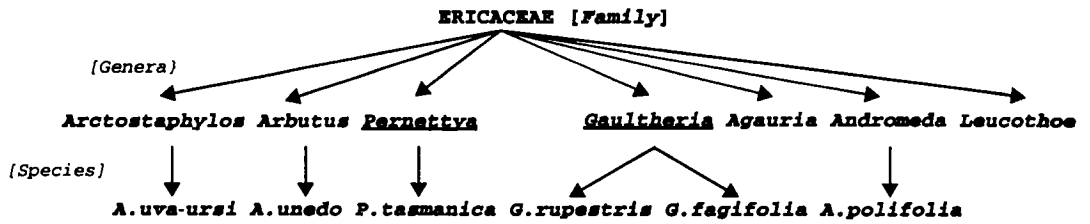

Figure 2: *Input hierarchy implemented in ReTAX for the revision of family Ericaceae*

As a consequence, the system applied OP.6, and renamed all the species of *Pernettya* as species of *Gaultheria.*

The results of ReTAX's implementation are consistent with the historical changes that led to the evolution of the taxonomic status of *Pernettya* and *Gaultheria,* as described in Section 3.2.

4 Final Remarks

We have presented a model for scientific taxonomy revision which takes into account psychological evidence of categorisation by expert biologists, and historical evidence from the domain of botanical classification. The combination of inductive procedures (involving the comparison of items) with taxonomic knowledge about the discriminant value of the descriptors, together with the application of other "feature-updating" and "hierarchy re-structuring" operators, have been used to replicate taxonomic refinements performed historically by plant taxonomists.

ReTAX has been tested so far with a limited subset of the taxa belonging to Ericaceae, and without considering other taxonomic groupings related to the family. A more realistic account of the task of taxonomy revision should involve the utilisation of larger knowledge bases with a larger number of taxonomic characteristics. We are currently working on the extension of the hierarchy we give to the system as input, in order to include a further number of taxa and consider a larger set of discriminant and non discriminant descriptors.

Some other extensions and enhancements we are planning to incorporate to ReTAX include:

- the application of the system to a different taxonomic domain (e.g., zoology, geology) to test the generality of the procedures implemented;
- the extension of the conditions under which ReTAX generates a refinement; so that it attempts the revision of a hierarchy not only as a consequence of finding novel inconsistent items but also as a result of the acquisition of new taxonomic information (e.g., the discovery of new biochemical knowledge);
- the utilisation of statistical analyses to aid in the revision of large populations of items;
- the use of ReTAX as a "student modelling system", to see if the refinement operators can reproduce the faulty or incomplete knowledge of a novice scientist.

We hope that the use of a system like ReTAX (together with other computational tools for biological classification like the ones referenced in Section 1) will aid botanists perform, in a more systematic and reliable way, the demanding and time consuming task of taxonomy revision. Finally, we plan to use the system as a tool to investigate, in more depth, the role that the revision of classification schemes can play in other aspects of the scientific work, such as experimentation, hypothesis formation, and theory revision.

5 References

Abbot, L. A., Bisby, F. A., Rogers, D. J. (1985): *Taxonomic Analysis in Biology.* New York, NY: Columbia University Press.

Alberdi, E & Sleeman, D. H. (to appear): Taxonomy revision as shift of representational focus. In *Proceedings of the European Conference on Cognitive Science (ECCS-95).* France, INRIA.

Davis, P. H. & Heywood, V. H. (1963): *Principles of Angiosperm Taxonomy.* Edinburgh, UK: Oliver & Boyd.

Domingo, M. (1993): Towards a knowledge level analysis of classification in biological domains. In N. Piera Carreté and M. G. Singh (Eds.): *Qualitative Reasoning and Decision Technologies.* Barcelona, Spain: CIMNE.

Ericsson, K. A. & Simon, H. A. (1984): Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Hooker, J. D. (1876): [Ericaceae]. In G. Bentham & J. D. Hooker: *Genera Plantarum,* vol II, Part 2. London, UK: Reeve & Co.

Karp, P. D. (1993): Design methods for scientific hypothesis formation and their application to molecular biology. *Machine Learning, 12,* 89-116.

Klahr, D., Dunbar, K, & Fay, A. L. (1990): Designing

good experiments to test bad hypotheses. In J. Shrager & P. Langley (Eds.): *Computational Models of Scientific Discovery and Theory Formation.* San Mateo, CA: Morgan Kaufmann.

Korpi, M. (1988): *Making conceptual connections: an investigation of cognitive strategies and heuristics for inductive categorisation with natural concepts.* PhD Dissertation, Stanford University.

Kulkarni, D. & Simon, H. A. (1988): The process of scientific discovery: the strategy of experimentation. *Cognitive Science, 12,* 139-176.

Lakatos, I. (1976): *Proofs and Refutations.* Cambridge, UK: Cambridge University Press.

Langley, P., Zytkow, J., Bradshaw, G., & Simon, H. A. (1983): Three facets of scientific discovery. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence.* Karlsruhe, West Germany: Morgan Kaufmann.

Middleton, D. J. & Wilcock, C. C. (1990): A critical examination of the status of *Pernettya* as a genus distinct from *Gaultheria. Edinburgh Journal of Botany, 47,* 291-301.

Nordhausen, B. & Langley, P. (1990): An integrated approach to empirical discovery. In J. Shrager & P. Langley (Eds.): *Computational Models of scientific discovery and theory formation.* San Mateo, CA: Morgan Kaufman.

Pankhurst, R. J. (1975): *Biological Identification with Computers.* London, UK: Academic Press.

Rajamoney, S. A. (1990): A computational approach to theory revision. In J. Shrager & P. Langley (Eds.): *Computational Models of Scientific Discovery and Theory Formation* San Mateo, CA: Morgan Kaufman.

Rose, D. (1989): Using domain knowledge to aid scientific theory revision. In *Proceedings of the Sixth International Workshop on Machine Learning.* Ithaca, NY: Morgan Kaufmann.

Sokal, R. R. (1974): Classification: purposes, principles, progress, prospects. *Science, 185,* 1115-1123.

Watson, L., Dallwitz, M. J., Gibbs, A. J., and Pankhurst, R. J. (1988): Automated taxonomic descriptions. In D. L. Hawksworth, D. L. (Ed.): *Prospects in Systematics.* Oxford, UK: Clarendon Press.

Woolley, J. B. & Stone, N. D. (1987): Application of artificial intelligence to systematics: SYSTEX - A prototype expert system for species identification. *Systematic Zoology, 36,* 248-267.