Word Formation and the Validation of Lexical Resources

Pius ten Hacken

Universit t Basel
WWZ — Abt. Geisteswissenschaftliche Informatik
Petersgraben 51, 4051 Basel, Switzerland
pius.tenhacken@unibas.ch

Abstract

In the framework of Word Manager (WM), morphological dictionaries are produced by the classification of lexemes in terms of a rule database. The intricate structure of the resulting lexical resources, conceived primarily for flexible use, also offers novel opportunities for the validation of the lexical specification. Many of the inconsistencies and errors encountered in lexical specification in a text file are excluded in WM, because the lexicographerÕs interface supports decisions by the exploitation of the procedural nature of inflection and word formation rules. There remains a set of lexicographic decisions, based on facts of the language and on the theoretical analysis of these facts, which cannot be supported in this formal way. They include the contents of the lexicographic guidelines. For the validation of these decisions, two types of browser are provided, the tree browser which gives access to partitionings of the set of lexemes, and the lexeme browser which concentrates on information for a single lexeme and on its links to other lexemes. The possibilities available because of the structure in the database constitute a challenge for the generality of the approach to validation described by Underwood & Navarretta (1997), which requires the reduction of lexical databases to text files.

1. Introduction

Two essential issues for the usability of lexical resources are the validation of the correctness and consistency of the information provided and the documentation of the treatment of particular phenomena. A standard description of a validation procedure is given in Underwood & Navarretta (1997). The procedure they describe is meant to be applicable to the most common type of lexicons for computational linguistics, consisting of text files listing entries with a variety of information types. In validation there is a certain trade-off between generality and precision. In this paper I will argue that, in a more limited context, word formation can be exploited so as to facilitate validation and raise the standard of quality and documentation which can be expected in the lexical resources concerned.

As a starting point I will take lexical resources for English and Italian currently being developed in Basel and Lugano in the framework of Word Manager (WM). In section 2, the basic features of the WM framework and the current project are described. Section 3 describes the validation process. Section 4 briefly addresses a related question, namely the usability of the resources developed in different theoretical frameworks. Finally, section 5 summarizes the main conclusions.

2. Context

Word Manager (WM) is a system for morphological dictionaries. It is one of the systems which emerged as a reaction to the ÔlexicalbottleneckÕproblem in the late 1980s. Contrary to most other approaches, it defines its scope not in terms of a division between lexicon and rule components, but in terms of the type of knowledge involved. A WM dictionary includes at the same time entries with morphological knowledge and operational morphological rules. Ten Hacken & Domenig (1996) offer a general description of the WM approach. Ten Hacken (1999) focuses on the contrast with the more traditional approach to lexical resources.

The development of lexical resources in WM can be divided into three stages. First, a rule database is specified, describing the morphological system of a language. This description includes the inflection rules and the word formation rules of the language, each rule illustrated by at least one example. For irregular inflection rules, all entries are specified. In the description of word formation rules, all affixes are specified. In the second stage, lexical entries are entered by attributing them to the classes defined in the rule database. From this morphological dictionary, dedicated tools are derived in the third stage. For each stage, a special user interface has been created supporting the particular task to be performed.

In the course of the project ÒWord Formation as a Structuring Device of the English and Italian Lexicons: A Large-Scale ExplorationÓ WM lexical resources for English and Italian are being developed by small teams of two lexicographers each, based at the Universit t Basel and the Universit della Svizzera Italiana (Lugano). As a starting point three types of constraints are available for guiding the specification: the WM environment, including its formalism, its compiler, and its interfaces; the rule databases for English and Italian; and lexicographic guidelines on the analysis of problematic or controversial phenomena.

The most prominent aim of the project is of course to produce lexical resources in the WM format for both languages by classifying entries in terms of the rules. A simple entry, i.e. a lexeme not resulting from a word formation rule, is assigned to an inflection class (IRule). The IRule generates the inflectional paradigm. A complex entry is assigned to a word formation rule (WFRule). The WFRule models the underlying word formation process and assigns the resulting entry to an IRule for generating the inflectional paradigm. The specification process is supported by a menu which prompts the lexicographer to select a WFRule, find the appropriate base lexeme(s) in the database, and select any affixes involved in the process. On the basis of this information, the system generates the resulting lexeme.

As a side effect of this classification task, problematic phenomena for the rules or the guidelines are discovered. Two types of cases can be distinguished here. First there are lexemes which cannot be classified because an element for their specification is missing, e.g. an affix not available in the rule database. Second there are lexemes for which it is not clear which classification should be chosen. As part of the project, a procedure for the systematic collection of such problem cases has been devised. A second aim of the project is then the improvement of the rule databases and the guidelines on the basis of the analysis of the these problem cases.

The output of the project, which runs from July 2000 to June 2002, consists of morphological dictionary databases of approximately 60,000 entries each for English and Italian and a set of lexicographic guidelines which can serve as documentation of these databases. As a general policy, the guidelines are common to both languages. Exceptions are made only for obviously language-specific phenomena, e.g. the distinction between British and American English. The information available for lexemes is very rich at morphological level but excludes other levels, e.g. subcategorization or semantics. This is a conscious choice inherent in the WM approach. The implications of this decision are discussed in section˚4.

3. Validation

The conditions to be complied with in the specification of lexemes in WM can be divided into four types according to their source:

(1) The WM environment
(2) The morphological rule database
(3) The lexicographic guidelines
(4) The linguistic facts

In the validation of WM resources it should be checked that (4) is described while observing (1-3). From the point of view of the lexicographer making decisions about how to specify a particular lexeme, (1-2) are ÔhardÕ constraints which cannot be violated because the system does not allow violations, whereas (3-4) are ÔsoftÕ constraints, knowledge to be kept in mind during specification, but which the system cannot enforce. It is useful to consider the validation with respect to these two types of constraints separately.

3.1. Hard constraints

WM supports the specification of lexemes by providing a dedicated lexicographerÕs interface. The dedicated nature of the interface means that the information available in the database is presented in a format which supports the classification of lexemes in terms of IRules and WFRules, while hiding any information not relevant to this task. As a consequence, formal properties of the WM environment (formalism and compiler) are imposed on the lexicographic decision process in the same way as linguistic choices encoded in the rules. For changes in the morphological rules, a different interface is used.

From the point of view of validation, it is attractive to encode constraints as much as possible as hard constraints, because the impossibility of violating them in the lexicographerÕs interface guarantees consistency of the entries with the constraints. There are a number of properties of the WM formalism and the compiler which are relevant in this respect.

First of all, WM is a closed system in which every entity is explicitly declared. A lexeme is a list of word forms with a citation form, belonging to a particular IRule. A word form consists of a sequence of formatives. A formative is a (possibly empty) string of characters with a (possibly empty) set of features. Legal characters and legal features are explicitly listed in the database. Spelling rules (SRules) may change the string of a formative in a way comparable to two-level rules, but every variant of the string has to be listed in the specification of the formative. In the compilation of a rule database, each IRule and WFRule is applied to example entries so as to check that these constraints are observed. A more complete description of the formalism of WM can be found in Domenig & ten Hacken (1992).

A second class of constraints consists of feature dependencies. In the Italian database, it is specified, for instance, that each noun must have a gender feature, so that it is impossible to enter a noun without gender.

A third type of constraints concerns the availability of IRules for new entries. IRules are specified as regular or irregular (RIRules and IIRules). For IIRules, all simple entries are specified in the rule database. In the lexicographerÕs interface, only RIRules are offered as possible options. Thus the lexicographer cannot add irregular verbs unless they are complex and based on the application of a WFRule to an existing irregular verb.

In the lexicographic specification phase, the support provided can be summarized as follows. For the specification of a simple entry, a list of RIRules is offered. The lexicographer selects an RIRule, enters the lexical form and the surface form(s) and, if desired, additional features not provided by the IRule. WM calculates whether all constraints are satisfied and, if so, gives an overview of the lexeme produced in the lexeme browser. The lexicographer can accept it or return to the specification interface. An example of a constraint violation is when an SRule applies and produces a surface form not specified by the lexicographer, e.g. "compani" for *company* because of the regular plural *companies*. In such a case the system asks permission to add the missing surface form. For individual exceptions, entry-specific SRules can be added and non-existing inflectional forms can be deleted (e.g. first and second person forms of weather verbs).

For the specification of a complex entry, a list of WFRules is offered. When a WFRule is selected, the system knows how many stems and affixes it requires. For stems, the lexicographer is prompted to search for relevant lexemes in the IRule classes. A string match search mechanism is offered to support this task. For the selection of affixes, the list of possible affixes associated with the selected WFRule is presented in a pull-down menu. When the source formatives have been selected, the system generates the corresponding entry, to be inspected in the lexeme browser in the same way as for simple entries.

In this way, many general linguistic constraints cannot be violated by the lexicographer because they are included in the rule database. Lexemes of a type not foreseen in the

rule database cannot be encoded by the lexicographer. They require the revision of the rule database in a different interface or a revision of the linguistic analysis. Changes in the rule database can only take effect after compilation. In compilation, all formal constraints are checked again. The lexicographer's entries are deleted in compilation. Therefore they have to be exported to a text file before compilation and imported after compilation. Only entries for which the rules are still available and the formatives have not changed can be imported. Other entries are listed in a log file and have to be entered again, so as to avoid inconsistencies appearing by modification of the underlying rule database.

3.2. Soft Constraints

As constraints modelled in terms of the WM application and the rule database cannot be violated by the lexicographer, optimal support of the lexicographic specification demands that as much information as possible is represented in this way. In the project for English and Italian, the basic policy has therefore been to restrict the number of descriptive options available to the lexicographer.

At the same time it must be recognized that there are many decisions which cannot be formalized in terms of hard constraints. Many of these decisions can be summarized under the heading of "facts of the language". The Italian lexicographer simply has to know that *agenda* ("diary") is a feminine noun with the plural *agende*, whereas *aroma* ("fragrance") is a masculine noun which has *aromi* as its plural. This is of course knowledge any speaker of Italian will have, and if mistakes occur in such cases they are due to performance errors during the specification process. While the system can impose the assignment of gender and inflection class to nouns, it cannot impose in the same way the specification of the *correct* gender and inflection class to each noun.

The most problematic decisions are theoretical decisions not directly grounded in the language competence, which cannot be supported by the system. Whether *considerable* is analysed as a deverbal adjective based on *consider* (which subsequently specialized its meaning) or as a simple adjective not related to the verb is not a question of language competence, but a theoretical decision. What WM can do (and does) is signal a problem if both analyses are entered. It is not possible, however, to exclude one of the two options without recourse to *ad hoc* measures, because both analyses represent a large class of attested cases.

For such cases lexicographic guidelines are necessary. At the start of the project it was decided to adopt a uniform set of coding guidelines for English and Italian in order to increase the theoretical impact. Common coding guidelines have been developed for issues such as the distinction between etymology and word formation (cf. Ten Hacken & Smyk, 2002), the interaction of word formation and homonymy, and the treatment of neo-classical word formation (cf. Petropoulou & ten Hacken, 2002). Another type of problem for which guidelines were necessary concerns phenomena on which morphological theories are generally silent, e.g. abbreviations and proper names. When the lexicographers have fully understood and internalized these, they will make sure that the resulting database has the same degree of consistency in these cases as in contrasts such as *agenda* vs. *aroma*. At the same time, the guidelines are useful as documentation of the content of the database.

One of the by now generally accepted conclusions of discussions in philosophy of science is that observation is necessarily theory-dependent, cf. Margolis (1993). Even for objects in the physical world, the way they are observed, described, and classified depends on the structure of knowledge in the observer's mind. Otherwise there would simply be too much to observe about an object to have any hope of finishing the observation process. This is all the more valid for a mental device such as knowledge of language. Speakers of a language have to be taught to see certain analyses. The word formation analyses encoded in WM are not directly necessary to use the language, so that linguistic competence has to be supplemented by theoretical knowledge for the specification of lexemes in a WM database. As a consequence, there is no clear dividing line between the facts of the language and the theoretical decisions to be encoded in guidelines. What should be covered by the guidelines is ultimately an empirical matter: whatever turns out to be controversial in a given case.

Given these considerations, it is not possible to impose a rigid distinction of soft constraints in violations of (3) and (4) above. Rather, for a clear understanding of the validation process, they should be classified in terms of the following types:

(5) Performance-type errors
(6) Errors due to misunderstanding of the guidelines
(7) Inconsistencies due to missing guidelines

In (5-7) no *a priori* distinction is made between facts and theoretical decisions. Errors of type (5) are unsystematic divergences from what the lexicographer actually knows. Errors of the types (6-7) are typically much more systematic, because they are linked to knowledge, although not of the intended type. In the case of (7), the knowledge is typically not very explicit. As lexicographers are encouraged to report cases where they feel uncertain about the correct specification of an entry, (7) concerns cases where different lexicographers do not realize that there is a possibly controversial issue, but have conflicting solutions in their implicit theories.

3.3. Tools for the Validation of Soft Constraints

In the WM interface for lexicographic specification two browsers are integrated which facilitate the systematic exploration of the lexicon database. Of these two, the tree browser enables the user to define subsets of the database on the basis of rules, features, and affixes, and the lexeme browser gives access to different aspects of the knowledge associated with a particular lexeme.

In the tree browser, the lexicon database is considered as a set of entries. This set can be subdivided according to four criteria:

- Word formation rules
- Inflection rules
- Entry features
- Word formation formatives

In the rule database, WFRules and IRules are organized in a tree. At the highest level, word formation in the Italian database is divided into derivation, compounding, and neo-classical word formation. Derivation rules are divided into classes marked by the syntactic category of their input and output, and so on until the level of the individual WFRule. In the tree browser, we find this organization reflected in the classes of word formation as in Fig. 1.

Fig.˚1 shows the Italian database as it was in January 2002. In the word formation perspective shown here, the WFRules are considered as classes consisting of the entries formed by their application. Each line in the window refers to a subset of the database determined by a WFRule or a set of WFRules corresponding to a node in the WFRule hierarchy, and gives the number of lexemes in this subset. The highest-level division partitions the database into simple entries, not formed by a WFRule, and complex entries, resulting from the application of a WFRule. The triangle at the start of each line gives three types of information. First, its position with respect to the left margin of the window indicates the level of embedding of the class in the hierarchy of WFRules. Thus, *Compounding* is a sister of *Derivation*. Second, whether the class is currently collapsed or expanded is indicated by the orientation of the triangle. Thus, lines 2 and 5 are collapsed, lines 1, 3, and 4 expanded. Third, the existence of any further subclasses is indicated by filling or not. Thus, *Compounding* can be expanded into rules, but *Not WFRules* only into entries. For each class, the full set of entries can be retrieved in a separate window.

The other views listed can be triggered by the buttons on the top right of Fig.˚1. The inflection view of the tree browser is organized parallel to the view in Fig.˚1, reflecting the IRule hierarchy in the database. Entry features are all features associated with entire lexemes (as opposed to individual word forms). In the entry features view, the root is expanded into all attributes and each attribute can be expanded into its value set. Word formation formatives include all affixes. In the word formation formative view, the root is expanded into all shapes of affixes. Homonymous affixes can be subsequently expanded into the individual formatives with different feature assignments.

The power of the tree browser is increased significantly by the possibility of linking different views. By selecting a particular class in one view and choosing one of the buttons on the right, the second view is given for the subset selected in the first window. This is illustrated in Fig.˚2.

.

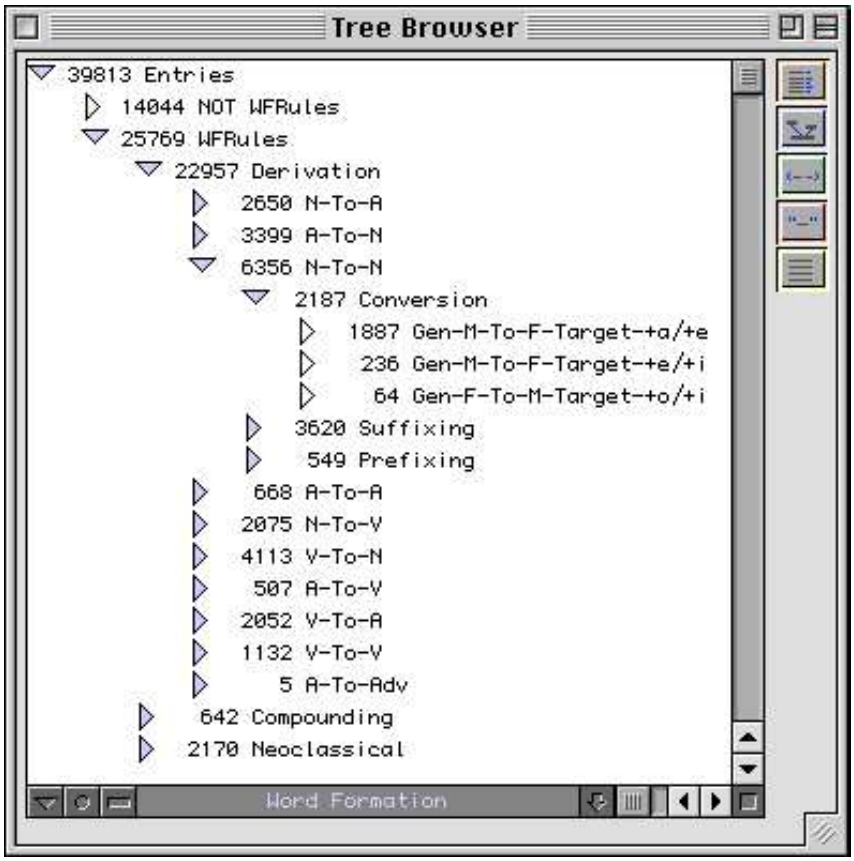

Fig. 1: Partially expanded word formation view of the tree browser for the Italian WM lexicon database.

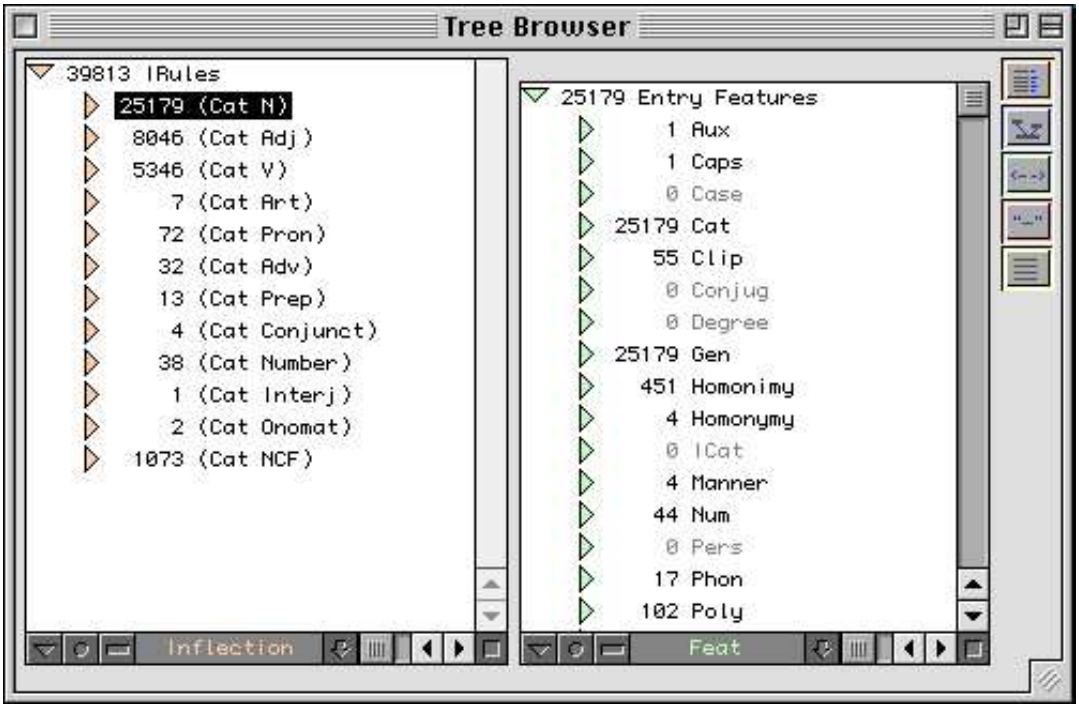

Fig. 2: Linked inflection and entry feature views.

In Fig.˚2, the subtree of the nouns is selected in the inflection view and linked to the entry feature view. The link is shown by the top margin of the right-hand window. As indicated by the scroll bar, the feature list continues further down. Fig.˚2 illustrates in two ways how the tree browser can be used in the validation of lexicographic specification. The first is exemplified by the two lines with *451 Homonimy* and *4 Homonymy*. The attribute in question is entered by the lexicographers for certain lexemes according to particular guidelines which we will not deal with here. The problem in this case is that there are different spellings for the same attribute. The English spelling is only used for four hard-coded entries. *Homonimy* constitutes a trivial case of (6) above, a systematic divergence from the guidelines, in this case a spelling error no doubt under the influence of Italian *omonimia*. We will come back to the possibilities of patching up such an error in section 4. The second type of error which can be detected in this way is indicated by the second line of the right-hand window. The feature *Aux* is used for verbs and indicates which auxiliary they have to form their perfect. In this case, it was accidentally entered also for one noun. This is a clear case of an unsystematic error of type (5), which can be corrected straightforwardly by deleting the Aux feature for the noun in question.

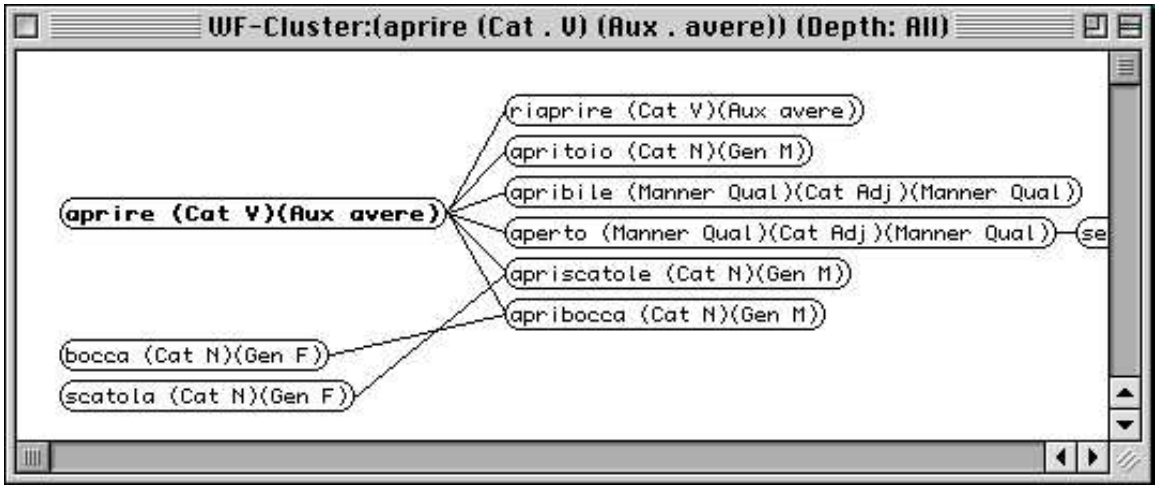

Fig. 3: WF-Cluster for the Italian verb *aprire* (ôpenõ), showing four derivations and two compounds. The adjective *semiaperto* is partially visible on the right-hand side.

As opposed to the tree browser, the lexeme browser focuses on a single lexeme. It can be opened by selecting a lexeme in the tree browser or in a text window. For inflection, it can show the IRule, the word forms, and the formation processes for the individual word forms. For word formation it can show the creation history with WFRule, formatives, and SRules involved and the generation history, a list of lexemes based on the lexeme in question. An overview of the word formation links is shown in the WF-cluster, illustrated in Fig. 3. Throughout the lexeme browser, all references to other entries are hot links to the corresponding lexeme browsers.

3.4. From Guidelines to Browsers

In the presentation of the browsers, some simple cases of their use in the discovery of anomalies in the lexicographic specification were included. An important advantage of the tree browser is the possibility of seeing quantitative correlations between different classes.

The examples in the previous section illustrated the use of linking two ways of partitioning the database, answering questions such as *Which features cooccur with nouns and how often?* A somewhat more systematic application of this facility starts with the identification of potential problems in the guidelines. An example of this type is the interpretation of the guidelines for neo-classical word formation. Neo-classical word formation encompasses the processes forming such words as *anthropology*, *anthropomorphous*, and *morphology*. The treatment adopted in the project, described by Petropoulou & ten Hacken (2002), assumes that items such as *anthropo* and *morpho* are neo-classical formatives. They have no syntactic category so that they have to be involved in morphological processes in order to get one. This is encoded in WM by assigning them the feature (Cat NCF) and marking them as fictional entries. In order to mark a lexeme as a fictional entry, the lexicographer has to tick a field in the specification dialogue. In the tree browser, fictional entries are marked with an *f* before their string. Fig. 4 shows how entries where this was forgotten can be identified. In the left-hand window, (Cat NCF) is selected as an entry feature. In the right-hand window, the list of entries is ordered alphabetically and a button makes it possible to select all fictional entries, which in this case highlights the incidental non-fictional ones.

This example is typical in the sense that it shows how the validation in terms of the guidelines requires first of all a hypothesis on possible errors which can then be translated into browser selections.

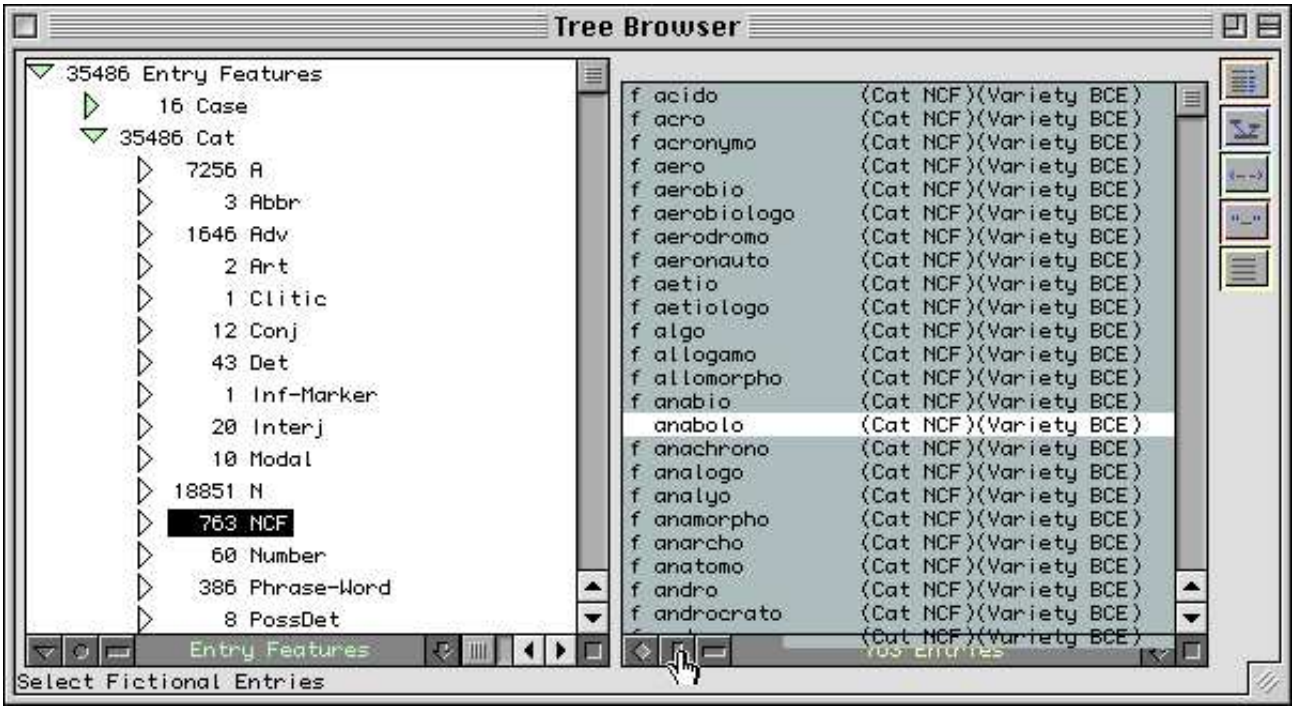

Fig. 4: Tree browser for the English database with selected fictional entries of (Cat NCF).

3.5. The Role of Word Formation

The development of lexical resources in WM is centred around morphology. A WM lexicon database is tightly structured by its use of inflection and word formation. It is this organization which makes the validation in terms of hard constraints as discussed in section 3.1 possible. The availability of the information about word formation considerably enriches the browsing possibilities used in the validation in terms of soft constraints.

A decisive property of WM, which distinguishes its resources from the more standard type, is the integration of declarative and procedural perspectives on rules. Ten Hacken (1998) shows how this property increases usability in a number of practical contexts. The validation procedures discussed here would not be possible without this integration.

A comparison of the validation of WM resources as discussed here and the general validation procedure described by Underwood & Navarretta (1997) shows one immediate conflict. Underwood & Navarretta assume that, as a preliminary step towards the validation of lexical resources, any database structure used in their development should be removed to produce a text file. In a text file, however, the interplay of different rule perspectives in WM would be destroyed. Of course it is possible to reduce WM databases to text files, in fact this is done in the export procedure described in section 3.1. The export files produced, however, are only meant to be used by WM in the import process. For human inspection of WM resources, it is much more sensible to produce a dedicated view of the database. The tree browser constitutes one such view which is currently available. Additional functions or alternative presentations of the information dedicated to the validation in terms of the guidelines, which would make particular types of queries possible, can be implemented by the use of the tool generator (cf. below).

Turning to the comparison of the actual procedures proposed for validation by Underwood & Navarretta (1997) and for WM, we find an important difference in the level of automatic support. Apart from relatively trivial questions such as whether all attributes and values in the feature declaration are actually used, the main task in Underwood & Navarretta's conception of validation consists of the manual inspection of a certain portion of the lexicon. This inspection takes the form of a rehearsal of the lexicographic decisions for the individual entries taken in the specification. In WM much more flexibility is offered, so that manual inspection can shift to a more global perspective of classes of entries. It is a well-known general fact about validation that its results improve when the data are approached from different perspectives. Therefore, while Underwood & Navarretta's approach is useful in many cases, the example of WM constitutes a challenge to its general applicability.

4. Usability

An aspect we have not addressed here so far concerns the reusability or indeed usability in practice of WM-based resources. The most common approach to the lexical bottleneck problem which emerged in the 1980s aimed to produce lexical resources independent of any particular application or linguistic theory. This is the general spirit evident in collections such as Atkins & Zampolli (1994) and Walker et al. (1994). A major problem for this approach is the existence and rapid development of a number of parallel, incompatible theories. The solution to this problem incorporated in WM links up, perhaps somewhat unexpectedly, with the issue of validation.

There are two main aspects to the WM approach to reusability. The first, discussed and justified in Ten Hacken (1999), is that WM takes as its domain not the lexicon as opposed to the rule component, but a linguistically determined domain (morphology) including entries and rules. The potential for inconsistencies due to theoretical discrepancies is reduced considerably if particular tasks are delegated rather than "the lexicon". For a neutral dictionary in the more common sense, theoretical choices in the rule component will interface with each individual entry of the lexicon. In a task-based approach, the interface with the client application concerns only the specification of the input and output of the task.

The second aspect of the WM approach to reusability concerns the flexibility of the presentation of information to client applications. As described by Pedrazzini (1999), a facility has been developed to derive lexical tools dedicated to the solution of a particular task. In this derivation step, it is possible to reorganize the classification of entries and the information presented about them, as long as the basic classes and information are available. This not only solves many theoretical divergences, but can also be used to minimize correction efforts if validation uncovers errors.

An example of a theoretical conflict is the attribution of syntactic category labels to minor categories. Especially in generative traditions, deadjectival adverbs are often considered as a type of adjectives, e.g. Larson (1987), and words such as *before* as prepositions which can optionally be intransitive or have a sentential complement. In the English and Italian lexicon databases, there are rules for deriving adverbs from adjectives and separate categories for prepositions, conjunctions, and adverbs to encode the different uses of items such as *before*. This does not mean that the lexical tool through which the database is accessed must have these categories. In the derivation of the tool, the classes can be redefined on the basis of the available information.

An example where this facility can be used to correct errors is one of the problems illustrated in Fig. 2 above. In principle, in order to correct an error, each entry has to be changed in a particular menu of the lexicographer's interface, which guarantees against the introduction of inconsistencies. In the case of the noun with Aux feature, this is not a particular problem. In the case of the misspelling of *homonymy*, it will be more practical to recode the four hard-coded entries and treat the change from *homonimy* to *homonymy* in the interface.

5. Conclusion

Lexical resources produced in the Word Manager system are highly structured. The structure is based to a significant degree on the representation of word formation relationships. The availability of this structure offers a

high degree of flexibility of access, which can be exploited both in the validation and in the practical use of the lexicons.

Compared to the validation procedure specified by Underwood & Navarretta (1997), a much more sophisticated set of tools is provided, so that it is no longer necessary to concentrate on redoing lexicographic work. A condition for the use of these tools is that the requirement that resources have to be presented in the form of text files is dropped.

Acknowledgements

The research project described here is financially supported by the Swiss National Science Foundation by grant 1214-058936.99. The lexicographic specification is carried out by Marco Passarotti, Evanthia Petropoulou, Chiara Restivo, and Dorota Smyk. Different types of support are provided by Antonia L scher, Bruno Moretti, and Sandro Pedrazzini.

References

Atkins, B.T.S. & Zampolli, A. (eds.) (1994). Computational Approaches to the Lexicon. Oxford: Clarendon.

Domenig, Marc & ten Hacken, Pius (1992). Word Manager: A System for Morphological Dictionaries. Hildesheim: Olms.

ten Hacken, Pius & Domenig, Marc (1996). Reusable Dictionaries for NLP: The Word Manager Approach. Lexicology 2, 232-255.

ten Hacken, Pius (1998). Word Formation in Electronic Dictionaries. Dictionaries 19, 158-187.

ten Hacken, Pius (1999). Two Perspectives on the Reusability of Lexical Resources. McGill Working Papers in Linguistics 14, 39-49.

ten Hacken, Pius & Smyk, Dorota (2002). Word Formation versus Etymology in Electronic Dictionaries. To appear in the proceedings of Euralex 2002, K¿benhavn, 13-17 August 2002.

Larson, Richard K. (1987). ÒMissing PrepositionsÓ and the Analysis of English Free Relative Clauses. Linguistic Inquiry 18, 239-266.

Margolis, Howard (1993). Paradigms and Barriers: How Habits of Mind Govern Scientific Beliefs. Chicago: University of Chicago Press.

Pedrazzini, Sandro (1999). The Finite State AutomataÕ Design Patterns. In Champarnaud, Jean-Marc; Maurel, Denis & Ziadi, Djelloud (eds.), Automata Implementation, Third International Workshop on Implementing Automata, WIAÕ8, Rouen, France (pp˚213-219). Berlin: Springer.

Petropoulou, Evanthia & ten Hacken, Pius (2002). Neoclassical word formation in an electronic dictionary. To appear in the proceedings of Euralex 2002, K¿benhavn, 13-17 August 2002.

Underwood, Nancy & Navarretta, Costanza (1997). A Draft Manual for the Validation of Lexica: Final Report. ELRA.

Walker, Donald E.; Zampolli, Antonio & Calzolari, Nicoletta (eds.) (1995). Automating the Lexicon: Research and Practice in a Multilingual Environment. Oxford: Oxford University Press.

.