Word Manager and Banking Terminology: Industrial Application of a General System

Daniela ZAPPATORE and Pius TEN HACKEN, Basel, Switzerland

Abstract

In practice, the use of specialized terminology in a firm's internal documents is often perceived as an obstacle to correct understanding of a text. With the emergence of intranet communication, the possibility arises of making available on-line explanations of terms used in a document. A condition for the success of such an enterprise is that terms are recognized also when they occur in inflected forms or when they consist of more than one word. Word Manager (WM) is a system for reusable morphological dictionaries which is optimally suited to these conditions. In addition it offers a user-friendly environment for the specification and maintenance of a database, and a tool generator which provides a model for the specification of a finite-state transducer mediating between the WM database and the terminology database. The collaboration between UBS AG and WM has shown that these advantages can also be realized in practice.

1 The Terminology Problem

Using a well-defined terminology can be a useful aid to specialized communication, but in less favourable circumstances it can also constitute an obstacle to understanding. This problem is highlighted in a situation where communication involves people used to similar but not identical terminological conventions. Since much of the prescriptive, norm-forming process involved in terminology is firm-specific, a merger is an occasion when terminology requires special attention.

In order to improve internal communication UBS AG have decided to facilitate access to terminological definitions by means of a tool which can be applied to documents on the intranet. Its design criteria are that for documents published on the intranet, employees can select a word or a string of words in a text and, if it constitutes a term, have the official explanation of the term displayed in a separate window. It is the intention that this service, called COFFEE (Conceptual Output Formatting For Easy Enquiry), should be available in German, English, French, Italian, and Spanish, but initially only German and English are considered.

One of the general problems of terminology is that terms behave not only as terms, but also as words and expressions of a language, taking part in syntactic and morphological processes. It is not enough to have the forms *option* and *oversubscribe* in the database, because we can also expect *options, oversubscribed* and *oversubscription* to occur in a text. Of course, this situation contributes to the tension between actual use and proper use, that is between descriptive and prescriptive applications of the term database, which is a well-known phenomenon in terminology in general (cf. Pearson, 1998:7-40).

The problem to be solved by the COFFEE service can be divided into three components. First, we need a term database, containing the information to be displayed for a given term. Second, we need a recognition module which reduces a selected word or string of words to its citation form. Finally, a match between the citation form and the term database entry has to be made.

2 The Word Manager Contribution

Word Manager (WM) is a system for reusable morphological dictionaries. Its general architecture and design objectives are described in Ten Hacken & Domenig (1996). In WM, the morphological rule system of a language, including its inflection and word formation rules, is specified in a morphological rule database. Lexical entries are specified in terms of these rules to constitute a morphological dictionary database. Integrated with WM is a component for the treatment of multi-word units, Phrase Manager (PM). As described by Pedrazzini & ten Hacken (1998), the morphological dictionary database can be used to derive fast-running finite-state transducers.

In the COFFEE service project, WM provides the recognition module and the transducer for mapping between the citation forms and the term database. WM was chosen because it has a number of advantages in these roles. First, it provides an integrated treatment of one-word and multi-word terms in a single database. This possibility is particularly important as for example a German compound word such as *Bundespräsident* may corresponds to a multi-word unit in English (*President of the Confederation*), French (*Président de la Confédération*) and Italian (*Presidente della Confederazione*). Second, terms are integrated with the morphological and phraseological rule mechanisms of the language. They are automatically part of classes determined by their phrasal, inflectional, and word formation properties. Third, a number of the specific modules required were already available. In particular, rule databases for German, English, and Italian existed as well as a lexical tool generator for the derivation of transducers. Here we will be mainly concerned with the specification of terminological entries in the English database.

3 The Specification Interface

The lexicographer has at his/her disposal an already existing and fully defined morphological rule database, which is general enough to accommodate the general word formation processes as well as the more specific ones that occur with a higher frequency in the formation of terms. The term database provided by UBS AG serves as the corpus, determining which entries should be made. All entries in the UBS term database have an index (number) and whenever possible the translation equivalents in English, German, Italian and French are given. The entries contain further information such as definition, notes, quotes, etc. All information supports the lexicographer's work, but only morphological information is entered in the WM terminological dictionary.

The existing lexicographer's interface in WM has been adapted for the work with terminological databases by the addition of the TD (Term Database) window and the TM (Term Manager) menu. The TD window gives access to the term database and can transfer selected forms directly into the environment for the specification of entries by the WM lexicographer. The term database is available as a document in a format similar to SGML, which contains the citation form in the languages for which it is available and the index. The TD window (Fig. 1) is the starting point for the lexicographer's work.

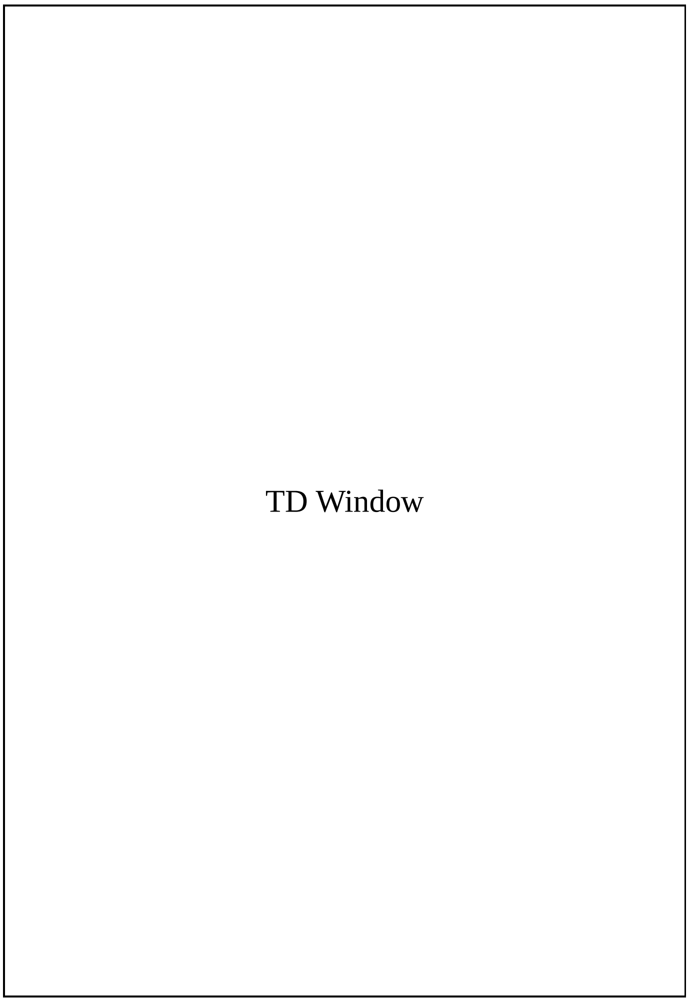

Figure 1: The TD (Term Database) window

The top part of the TD window in Fig. 1 offers ways to browse or navigate through the term database. The four buttons REW (rewind), NEXT, PREV (previous), and SEEK allow the lexicographer to move back and forth in the list of terms and to search it for a particular item (word or term) or index. The radio buttons on the right-hand side show the language chosen and give access to the corresponding terms in other languages. The remaining part of the window offers facilities for specifying words and terms. The dividing line in the middle of the window separates the two possibilities for making entries, namely words in the WM lexeme dictionary (upper part) and terms in the WM terminological dictionary (lower part). In the COFFEE service only the terms, i.e. the expressions contained in the term database, should be recognized, not the individual elements they consist of (in particular for multi-word terms). If an entry is specified via the buttons below the dividing line, this entry is automatically provided with an index, which marks it as a term. A more detailed description of the functioning of these buttons will follow in 4.1. and 4.2. The button ADD TERM is for multi-word entries, as will be discussed in 4.3. To its left is the UPDATE button, whose function and importance will become clear in 5.2. In the frame at the bottom of the window the term to be insertedappears in the citation form and above it the number with which it is indexed in the corpus.

4 The Lexicographer's Tasks

The treatment of a term by the lexicographer is determined by the fact that a term is a lexeme with a special marking. The lexeme has to exist before the corresponding term can be defined. The general architecture of WM as described in Ten Hacken & Domenig (1996) requires that more complex lexemes require the prior existence of the corresponding less complex lexemes. For the specification of a multi-word unit this means that the lexemes from which it takes its word forms should exist and a lexeme resulting from word formation can only be specified after the less complex lexemes it is built from.

4.1 Simple Entries

Let us start with an example of a one-word term, e.g. *intercompany* as illustrated in Fig. 1. Assuming that *intercompany* is not yet specified as a lexeme in the database, the lexicographer first has to enter it as such. The first decision concerns whether the lexeme in question is simple or complex. If the lexeme is the result of a word formation rule, the components used by the rule and the type of process involved (derivation, compounding, conversion) are specified. In the case of *intercompany*, the lexicographer will analyse it as a derivation involving the prefix *inter* and the stem *company*. Affixes such as *inter* have been specified in the database as part of the linguist's description of the rule system. Lexemes such as *company*, however, have to be entered by the lexicographer.

In order to enter *company* as the first step towards the specification of the term *intercompany* in Fig. 1, the lexicographer selects the string *company* in the box at the bottom of the TD window and clicks SIMPLE ENTRY in the upper part of the window. This opens the ADD SIMPLE ENTRIES window shown in Fig. 2. The string *company* is automatically shown as the value of "Corpus" and in the box in the middle of the window. The specification of an entry involves the listing of its lexical form and all its possible surface forms. The terms *lexical form* and *surface form*, explained in detail in Ten Hacken & Domenig (1996), are used in approximately the same sense as in Koskenniemi's (1983) two-level morphology. In the case of *company*, the surface form "compani" occurring in the plural has to be added by the lexicographer.

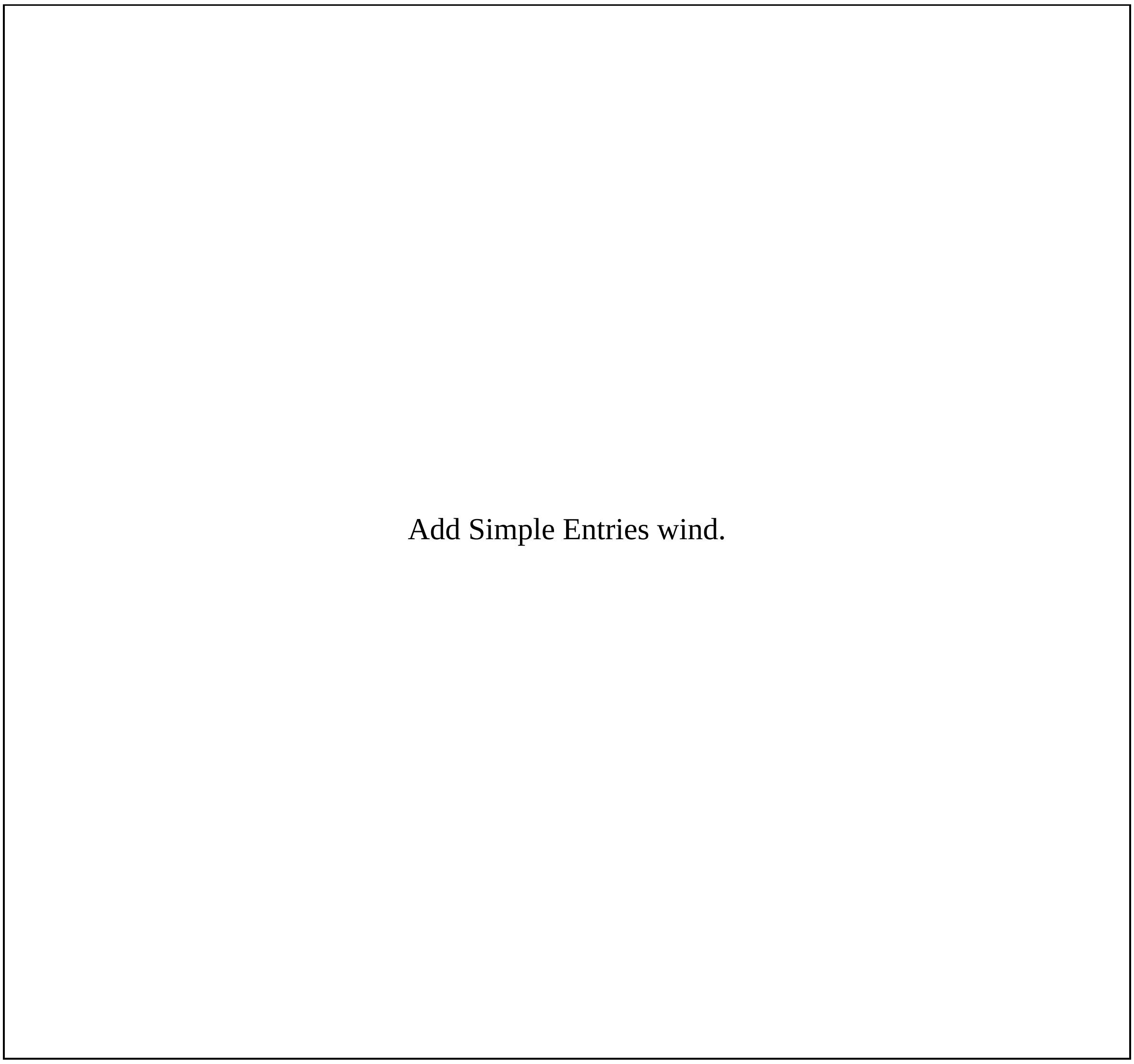

Figure 2: The Add Simple Entries window.

The specification of the lexeme *company* in Fig. 2 proceeds by the selection of an inflection rule from the listing in the top half. Other boxes are for the specification of additional information in the form of added features, deleted forms (e.g. the plural of mass nouns), or entry-specific spelling rules. Clicking GENERATE/ADD in the situation displayed in Fig. 2 results in the opening of the Virtual Entry Window in Fig. 3.

A virtual entry window shows different aspects of a lexeme, e.g. its word forms and the formatives they are built from. It is described in more detail by Ten Hacken (1998). By clicking ADD in Fig. 3 the lexeme is added to the database.

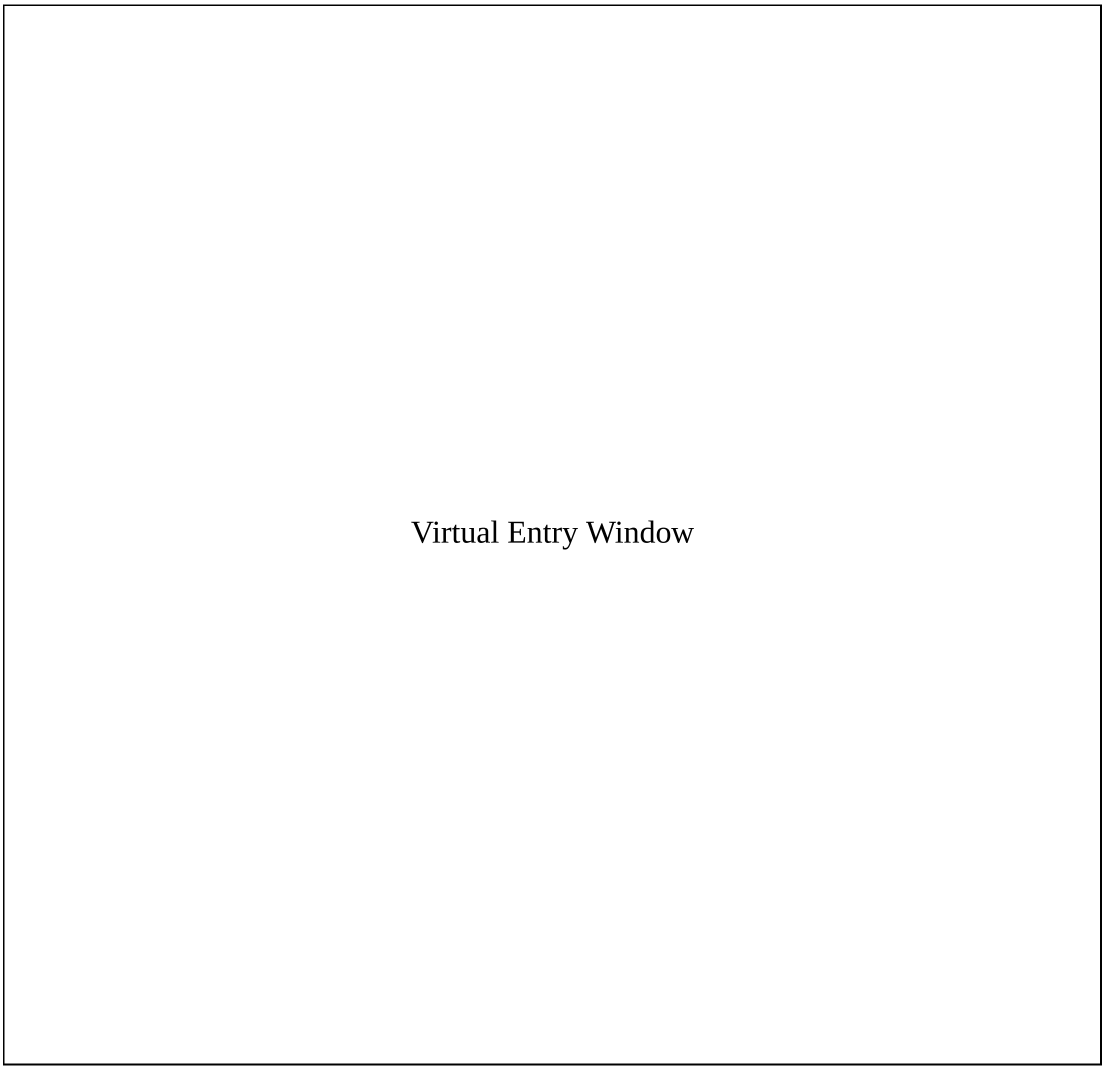

Figure 3: The Virtual Entry Window

4.2 Complex entries

The button COMPLEX ENTRY in the TD window submits the string selected in the box at the bottom to WM for analysis. As described by Ten Hacken (1998), word formation rules and formatives in the database are used to propose a structure for the new lexeme. The parse proposals are displayed as trees in the PARSES OF "intercompany" dialogue box (Fig. 4). The intermediate nodes in the tree indicate the word formation rule by a name. This name is a short form of the rule. Clicking on the name displays a bar with the expanded version of the rule. The top node indicates the inflectional rule (IRule). The lexicographer is asked to select the correct parse and another VIRTUAL ENTRY window is opened showing the consequences of this choice. At this point it can be accepted, rejected, or modified. The lexicographer can specify entry-specific word formation and/or inflectional spelling rules, and additional features, and deleted forms in the same way as for simple entries.

The Parse Window

Figure 4: The Parse Window

At this point, when *intercompany* has been specified in the WM database as a lexeme, the corresponding term can be inserted in the WM term database. The lexicographer selects the frame containing the term in the TD window (Fig. 1). The buttons below the dividing line in the middle are activated. The decision whether to enter the term AS COMPLEX TERM or AS SIMPLE TERM depends on the analysis of the underlying lexeme. In the case of *intercompany*, the former option is chosen. The only difference between an entry in the WM lexeme dictionary and one in the WM termdatabase is that the entries in the WM term database are provided with an index thatcorrelates them with the entries in the UBS term database. This index appears automatically in the Add Simple Entries window (Fig. 2) at the bottom of the window next to "Corpus Index".

4.3 Phrases

As mentioned above, the WM system has the advantage of offering an integrated module for multi-word units. The importance of such a possibility has been illustrated in section 2. The following types of information are needed in the specification of a multi-word term:

- HEADPHRASE, i.e. the citation form of the phrase to be specified.
- RESTRICTIONS: Restrictions define the word forms of the individual words that are allowed to occur in the specific context.
- MODIFICATIONS: In the context of PM, MODIFICATIONS do not refer to a syntactic relationship, but to the possibility of interrupting the linear sequence of the elements defined in the HEADPHRASE.
- CLASS: PM rules are organized according to phrasal classes (NP, AP, PP, and so on) and then further divided into subclasses. CLASS defines the rules with which the phrase is to be associated.

The individual lexemes that make up the multi-word term must be specified in WM beforehand.

The specification work of the lexicographer is supported in various ways by the system. In the context of the COFFEE project, multi-word units are only entered when they are terms, therefore – as already seen for WM term entries – the entire frame at the bottom of the TD window is selected. Recognition of multi-word terms in the TD-window is automatic: If the citation form in the frame contains a blank space, only the ADD TERM button below "ADD TERM (multiword or restrictions)" is activated. Clicking on it opens the ADD NEW PHRASE + INDEX window (Fig. 5). The upper box lists all the PM rules available to the lexicographer. The lower part is conceived as a fill-in slip with the four sections HEADPHRASE, RESTRICTIONS, MODIFICATIONS, and CLASS.

Add New Phrase-Window

Figure 5: The Add New Phrase-Window

With the opening of the window the citation form, *chief executive officer* in the example, is automatically inserted in the line below HEADPHRASE. The lexicographer only has to decide whether a multi-word term can be inflected. If this is the case s/he encloses each element that allows other inflected forms between angle brackets " <>". In the example only the noun *officer* can occur in the plural form. The exact definition of the word forms are computed automatically by the system when GET RESTRICTIONS is selected from the menu Assistant and inserted below RESTRICTIONS in the lower box. In case of homographs, however, the system asks for help by the lexicographer. A list of the homographs and their features is displayed so that the lexicographer can easily choose the correct lexeme.

In the next step the MODIFICATIONS are specified. MODIFICATIONS hardly ever apply to multi-word terms and are defined for an entry only when there is clear evidence for it in the corpus. A dash "–" in the line below MODIFICATIONS indicates that no interruptions of the linear sequence are allowed. Finally the lexicographer selects the appropriate syntactic rule from the list in the upper box. The name of the rule is also automatically shown under CLASS.

In the context of PM, the most striking difference between a general WM dictionary and a WM term dictionary lies in the set of classes defined in PM. The majority of multi-word terms are NPs, therefore the range of syntax trees can almost be restricted to the various types of NPs, leaving out specific VP-rules.

5 Special Problems

In the course of the projects various problems were encountered. We will focus here on two: capitalization and updates.

5.1 Capitalization

The tension between academic work and practical tools can be illustrated by the treatment of the use of capitals and lower case characters. In Domenig & ten Hacken (1992) it was decided that capitalization, as a primarily sentential phenomenon, was not a task of WM. In terminology, however, a variety of capitalization patterns is used, often distinctively. Furthermore, an exact match between the characters in a text and in the term database was a condition imposed by the UBS COFFEE project.

In principle there are two ways to introduce capitalization information into a system such as WM. At first sight, the most straightforward solution is to use upper case and lower case characters in the representation of the string. This solution guarantees optimal flexibility in the expression of idiosyncratic capitalization patterns, but it has a number of practical and conceptual disadvantages. A practical disadvantage is that updates are complicated, because WM does not foresee a change in the string of a lexeme. Conceptually, the character-based solution does not express the logic of the capitalization pattern. The second possible solution, which was in fact adopted, encodes the pattern in features assigned to the terms. In this way, common patterns can be encoded with simple features and exceptional ones stand out as such. One of the possibilities offered by WM is a "tree browser", showing a view of the lexical database as a class of lexemes which can be divided into subclasses according to various criteria, including individual features. The following cases are distinguished:

PATTERN	EXAMPLE	FEATURES
lower case	*option*	(Caps Normal)
first characters upper case	*Swiss Bank Corporation*	(Caps Wordcap)
all characters upper case	*OECD*	(Caps Upper)
first character of the components of a compound word upper case	*KeyPhone*	(Caps Special) (Spell KeyP)
in abbreviations, upper case characters determined by the written−out form	*BoD* (Board of Directors)	(Caps Special) (Spell BoD)

The features Caps, with four values, and Spell, with an open value set, are used. In this way, the regularity of the first three classes and the irregularity of the much smaller classes marked by (Caps Special) is expressed. Surface forms do not contain upper case characters and links between capitalized terms and corresponding general vocabulary items can be established by conversion-like word formation rules introducing the features Caps and Spell without affecting the string. The interpretation of the features Caps and Spell is the task of the transducer mediating between the WM-database and the term database.

5.2 Updates

For the maintenance of a terminology database, regular updates are necessary. They can be the result of changes in usage, new concepts, or the discovery of inconsistencies. Updates should be instantaneous from the point of view of the end user, which means that at no time should the service be down because an update is being made.

The update procedure starts from a periodic list of changes collected in an SGML file and made available to the lexicographer of the WM database through the TD window. The UPDATE button (cf. Fig. 1) opens a list of all existing entries corresponding to the term index of a particular entry from the update document. In this way, the decision to modify or add an entry is supported, so that updating does not create inconsistencies.

6 Conclusion

The UBS COFFEE project is very general in terms of its aims. It makes terminology accessible directly from intranet documents. In the realization of this aim, WM plays a mediating role between the text in the intranet document and the term database. WM recognizes the terms and their various inflectional realizations in the text and reduces them to their base form. This form can then be matched with the citation form as it is recorded in the term database and hence the definition and other information on the expression can be retrieved.

As a system for reusable morphological dictionaries, WM is ideally suited to recognize both single-word and multi-word terms in their morphological variations. Available resources which could be reused in the COFFEE project include the morphological rule databases and (for German) the full dictionary database. The entries for English specified in the course of the project are reusable in future applications. In addition, the transducer mediating between the WM output and the terminology database could be modelled after existing ones. In the development environment, only the TD window was added.

In the course of the project it turned out that the automatic representation of lexical and morphological relationships between terms in WM had the side effect of uncovering a number of inconsistencies and infelicities in the database, not immediately apparent to the human observer. The COFFEE project therefore also resulted in the improvement of the original term database.

At the time of writing, the German and English WM term dictionaries have been fully compiled and a start has been made with the Italian one.

References

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.

Koskenniemi, Kimmo (1983), *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics Publications No. 11.

Pearson, Jennifer (1998), *Terms in Context*, Amsterdam: Benjamins.

Pedrazzini, Sandro & ten Hacken, Pius (1998), “Centralized Lexeme Management and Distributed Dictionary Use in Word Manager”, in Schröder, Bernhard; Lenders, Winfried; Hess, Wolfgang & Portele, Thomas (eds.), *Computers, Linguistics and Phonetics between Language and Speech, Proceedings of the 4th Conference on NLP, Konvens’98, Bonn, Germany*, Frankfurt am Main: Lang, p. 365-370.