SANDRO PEDRAZZINI
IDSIA, Lugano, Switzerland
University of Basel, Basel, Switzerland-Suiza

# *Treating terms as idioms*

**Abstract**
It is recognized that electronic management of terminology spans a large range of activities. It is important to have easy access to well structured terminologies, but it is also relevant to model and recognize terms within a text. This allows, among others, multilingual cross references and the development of advanced indexing tools.
The aim of this paper is to explain how we model and treat terms to allow their recognition during a text analysis.

## 1. Introduction

In our work we start from an existent bank terminology database which collects more than 4,000 terms, mostly in 4, some in 5 languages, and we use it as source for a system designed to model and recognize idioms. I will not enter into the details of the database, because they are not relevant for the purpose of this article. Suffice it to say that you can have access to the data through the "canonical form" or through an unique index, which is shared among all translations of a given term.
We distinguish between *simple terms* (one word term) and *complex terms* (multi-words terms). The canonical form of a simple term is its citation form, whereas the canonical form of a complex term is a form in which no syntactic transformation has been applied.
A simple term is a word that can be found in a text in one of its different morphological forms.
A complex term is a combination of words, describing a concept, that can be found in a text in different morphological (for the single components) and syntactic forms.
To recognize a simple term we need a morphological analyzer, which can eventually allow the restriction of the inflection paradigm.
To recognize a complex term we need a higher level analyzer, able to recognize, through its morphological module, the different parts of the term, and then able to recognize the entire term, even outside its canonical form.

## 2. The Modelling System

The system used to model and recognize terms is Phrase Manager (Pedrazzini, 94), originally conceived to model idiomatic expressions, designed and implemented some years ago. It has been used for the last few years to model English idioms. The main characteristics of the system are its morphology-based module, Word Manager (Domenig and ten Hacken, 1992), which serves as main morphological dictionary, its lexicographer workbench, which allows the modelling of idioms, and its pattern-matching approach in recognizing idioms in the text.
The system allows the definition in form of «classes» of some general characteristics common to a certain number of idioms. The linguist must specify a classification. Each new entry will share some features with other entries of the same class.
Here are the characteristics that can be shared through a class association:

- The word category sequence of the canonical form. The class which represents an idiom like *kick the bucket* (to die), would have the sequence *V-Det-N*.

- The syntactic transformations that the elements of a class are able to undergo without losing their idiomatic meaning, i.e. the allowed permutations of the canonical word sequence. In fact most idioms exhibit certain transformational deficiencies (Swinney and Cutler, 1979; Cutler, 1982). The expression *kick the bucket,* for example, cannot be passivized. Idioms that do not support any transformation are *completely frozen*.

- All modifications, i.e. insertion of external elements, that are permitted for all idioms associated with the class. The former example would have no modifications (Cutler, 1982), but others, like *pull strings* (to exert secret influence) could be modified in different ways (*pull some illegal strings*) without losing their idiomatic meaning.

- Optional elements, i.e. elements that are parts of the idioms, but which absence would not change their idiomatic meaning.

Other characteristics or variations of these can be found in detail in Pedrazzini (1994).
Moreover each idiom can define different morphological restrictions for its single elements. In fact, the interaction between the phrasal system and the morphological system is fundamental. Words are the minimal components of the phrases and in Phrase Manager they can be divided into two categories: words that can inflect and words that cannot. Within words that can inflect, a further distinction can be made between those which may occur in all inflectional forms of their paradigm and those which may not.

There are some considerations that are common to idioms and to complex terms: first, flexible idioms (not completely frozen) appear in texts in different syntactic forms; thus the system must be able to describe these forms and associate them with their corresponding canonical forms; second, idioms can be divided into well-defined classes according to their syntactic structures and transformational degree.

The system must also be able to cater for these characteristics. The solution provided is based on a formalism which, first of all, allows the specification of classes which serve to represent syntactic behaviour which is common to entire sets of expressions and not merely individual ones. Individual expressions are inserted into the system as instances of classes and may therefore be specified only after a complete set of classes has been established.

## 3. The Recognition System

Our main interest is recognizing terms within a text. Terms can appear in different forms, depending on the characteristics associated during the definition.
The approach followed to recognize terms in a text is the pattern matching based on regular expressions.

We distinguish two levels: the morphological level and the word sequence level. For the morphological level we use an inflection transducer (Pedrazzini and Hoffmann, 1998), generated from our morphological module, able to analyze each single word and deliver the corresponding citation form. We show an example of such a transducer in Fig. 1, where the forms *dined*, *dine* and *dines* all lead to the citation form *dine*.

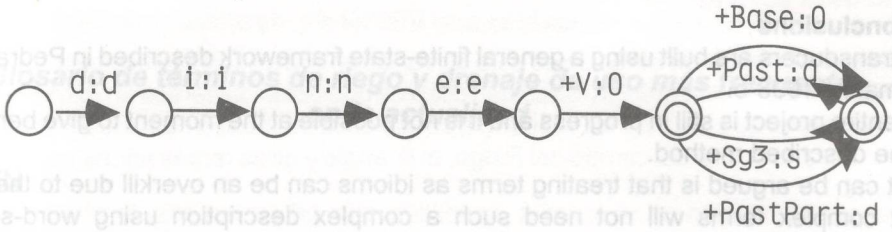


*Fig.1: Inflection Transducer*

For the word sequence level we build a second transducer containing information about the canonical form, all possible word sequences, modifications and transformations of a complex term. Moreover for each word, component of a term, it is possible to restrict the inflection paradigm or to define agreements with other components. We show the main components of the transducer in Fig.2, using the complex term open twin safe door.

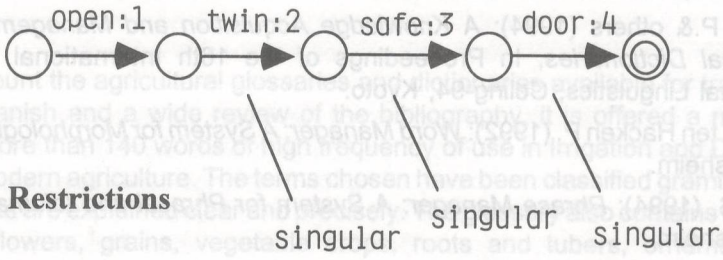


*Fig.2: Sequence transducer*

Information for both transducers are directly generated by the modelling system: morphological information by the morphological module (Word Manager), and terminological information by Phrase Manager.

An identical treatment is reserved to single terms. I said before that to recognize a simple term we only need a morphological analyzer. This is in general true. But in our system model, the inflection transducer is also meant as the basis dictionary for the whole system. Because simple terms may have a restricted paradigm, it would not be a good solution to restrict the paradigm of elements of the basis dictionary, which may need their entire inflection paradigm in other (complex) terms. Examples of simple terms with restricted paradigm in our term database are *Edelmetalle*, the German translation of the English complex term *precious metals*, and *Treuhandgeschäfte* (*fiduciary transactions*); both are treated as terms only in their plural forms.

For this reason we have decided to add simple terms to the sequence transducer. The will be defined using a single node, but they will have the opportunity to be restricted in their

inflection degree, without limiting the recognition power of the general inflection transducer. This design allows at the same time a clear distinction between term level (sequence transducer) and word level (inflection transducer).

## 4. Conclusions

The transducers are built using a general finite-state framework described in Pedrazzini and Hoffmann, 1998-b.

The entire project is still in progress and it is not possible at the moment to give benchmarks on the described method.

What can be argued is that treating terms as idioms can be an overkill due to the fact that most complex terms will not need such a complex description using word-sequence, transformations, modifications and word restrictions.

If this is true for many terms, above all nominal terms, it can be demonstrated that others, like *open twin safe door*, or even *offline table must be empty before branch is opened*, can be better retrieved in a text if their description covers their possible modifications and transformations, exactly as it may happen with idioms.

Because the recognition of terms in a text will be more and more relevant, not last to allow multilingual cross references and the development of advanced indexing tools, we think that sophisticated tools should be available to allow a precise and selected recognition of more and more complex terms.

## References


Cutler A. (1982): *Idioms: The Colder the Older,* Linguistic Inquiry 13, 317-320.

ten Hacken P.& others (1994): *A Knowledge Acquisition and Management System for Morphological Dictionaries*, In Proceedings of the 16th International Conference on Computational Linguistics, Coling-94, Kyoto.

Domenig M., ten Hacken P. (1992): *Word Manager: A System for Morphological Dictionaries*, Olms, Hildesheim.

Pedrazzini S. (1994): *Phrase Manager: A System for Phrasal and Idiomatic Dictionaries*, Olms, Hildesheim.

Pedrazzini S., Hoffmann M. (1998-a): *From Lexical Acquisition to Lexical Reusable Tools*, First International Conference on Language Resources and Evaluation, Granada, Spain, May 1998.

Pedrazzini S., Hoffmann M. (1998-b): *Using Genericity to Create Customizable Finite-State Tools*, FSMNLP'98, International Workshop on Finite State Methods in Natural Language Processing, Ankara, Turkey, June 1998.

Swiney D., Cutler A. (1978), *The access and processing of idiomatic expressions,* Journal of Verbal Learning and Verbal Behaviour 18, 523-534.