Terms Specification and Extraction within a Linguistic-Based Intranet Service

Sandro Pedrazzini, Elisabeth Maier, Dierk K nig

University of Basel
Klingelbergstrasse 50
4056 Basel (Switzerland)
and
SUPSI/IDSIA
Galleria 2
6928 Manno (Switzerland)
sandro@idsia.ch

Canoo Engineering AG
Kirschgartenstrasse 7
4051 Basel (Switzerland)
{elisabeth.maier,dierk.koenig}@canoo.com

Abstract

This paper describes the adaptation and extension of an existing morphological system, Word Manager, and its integration into an intranet service of a large international bank. The system includes a tool for the analysis and extraction of simple and complex terms. As a side-effect the procedure for the definition of new terms has been consolidated. The intranet service analyzes HTML pages on the fly, compares the results with the vocabulary of an inhouse terminological database (CLS-TDB) and generates hyperlinks in case matches have been found. Currently, the service handles terms in both German and English. The implementation of the service for Italian, French and Spanish is under way.

1. Introduction

In the last years intranets have been used as the main communication platform for large enterprises. At the same time, the number of documents posted on inhouse Web pages is growing rapidly. The amount of time employees spend for the search and processing of relevant, mission-critical information increases proportionally.

As one measure to reduce the time required for document a project was set up to tightly integrate an existing terminological database, CLS-TDB, into the intranet of the UBS AG, one of the world s largest banks. This is the reason why this article does not consider general extracting mechanisms, like the one described in Frantzi and Ananiadou (1996) or in Daille (1996). Our starting point is an existing and consolidated collection of terms.

The terminological database contains more than 4,000 terms, mostly in four, some in five languages. Before the new service was introduced data could only be consulted through a web-enabled database interface. Also, a list of all terms included in the CLS-TDB could be generated in the form of a text file. Now, terms defined in the CLS-TDB are highlighted in documents displayed in the browser, regardless of the morphological form in which they occur. Additionally, the definitions of these terms are accessible via hyperlinks, i.e. by clicking a highlighted term its corresponding CLS-TDB information is displayed.

As mentioned above, the system is able to identify all possible forms of both simple and complex terms. "Possible forms here refers to all paradigm word forms for simple i.e. one-word terms. For complex, i.e. multi-word terms the meaning is different. To recognize a complex term in a text, not only all allowed word forms for each single component have to be considered, but also all possible syntactic transformations and modifications, e.g. the insertion of external elements.

To recognize a simple term we need a morphological analyzer. To access database information for a simple term we need the citation form in one of the languages for which entries exist in the terminological database. The system component for the processing of simple terms consists of a morphological analyzer (implemented as transducer), which description can be found in Pedrazzini and Hoffmann (1998)

The recognition of a complex term requires a higher level analyzer; using its morphological module, this component allows to recognize the different parts of the term, and then able to identify the term in full, even if it is not occurring in its canonical form.

To access database information for a complex term we need its canonical form, i.e. a form in which no syntactic transformation has been applied. See Schenk (1989) and Verstraten (1989) for a more detailed description of a canonical form.

In a first step, the service was implemented for one-word terms. This paper describes how the service has been augmented in order to also handle complex terms, i.e. multi-word units.

In the following paragraphs the main development concepts and the run time behaviour of the system are explained.

2. The Treatment of Idioms and Complex Terms

The project started with the consideration that complex terms have a behavior similar or comparable to the behaviour of idiomatic expressions. Thus, a system conceived for the specification and recognition of idioms can also be used for terms.

The usefulness of systems for the recognition of idioms , e.g. in the field of machine translation, has not always been treated as an important matter: currently, machine translation systems are mostly used in very restricted fields, like e.g. in technical domains, where idiomatic expressions occurr only very rarely. This observation is also reported in Volk (1998). In the same paper it is also predicted that the need for idiom recognition is going to increase now that some machine translation systems are hooked up to the world wide web in order to translate arbitrary texts. Nevertheless, even if machine translation is used in very restricted domains where idioms are not used, it can be observed that exactly in these domains there is an extensive use of specialized terminology, often in form of complex terms. By exploiting the analogy between complex and idiomatic terms we can demonstrate the usefulness of systems for the recognition of idioms.

There are some considerations that are common to idioms and to complex terms:

¥ Flexible idioms, where the form is 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;

¥ Idioms can be modified and discontinuously spread inside a clause;

¥ Idioms can be divided into well-defined classes according to their syntactic structures and their transformational degree.

The system adopted to model and recognize complex terms is Phrase Manager (Pedrazzini, 1994). This system is based on Word Manager, which serves as morphological dictionary, and which has already been used for large dictionary implementations, as described in (Domenig and ten Hacken, 1996).

The work with Phrase Manager proceeds in two phases:

¥ *The modelling phase (Fig. 1)*
In a first stage the system is used to model and describe complex terms. It includes a user-friendly lexicographer workbench, which allows the specification of terms and the management of useful data needed for a pattern-matching approach, which is applied during analysis and extraction of idioms in a text. The solution provided is based on a formalism, which allows the specification of classes that describe the syntactic behaviour common to sets of expressions. Individual expressions are inserted into the system s database as instances of classes.

¥ *The generation phase (Fig. 2)*
On the basis of these data the system is used to compile and generate finite-state tools (FST). These tools are then used to extract terms from an HTML document.

In the following sections we give a detailed description of the different phases.

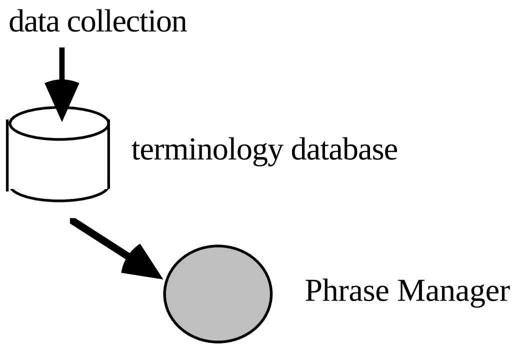


Fig.1: Modeling Data Flow

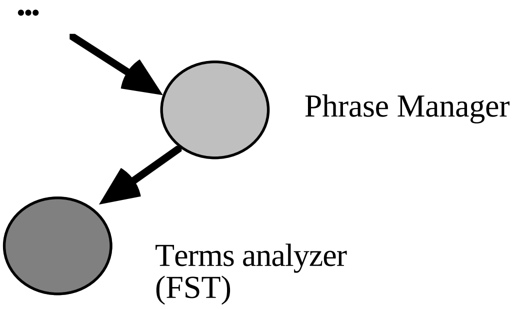


Fig.2: Generation Phase

3. The Modeling Phase

The system allows the definition of some general characteristics common to a certain number of terms. 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. We use idioms to illustrate the single characteristics of our classification:

¥ 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. According to (Swinney and Cutler, 1979; Cutler, 1982), most idioms exhibit certain transformational deficiencies. 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. While the former example would have no modifications (Cutler, 1982), others, like *pull strings* (to exert secret influence) can be modified in various ways (*pull some illegal strings*) without losing their idiomatic meaning.

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

Other characteristics are described in detail in Pedrazzini (1994).

For the modelling phase we used the inhouse terminological data base collected within the UBS AG. This data source contains a set of 4000 banking terms together with their definitions. Most terms are available in four, some in five languages.

In a first step, the German entries were modelled; the treatment of the English terms is close to completion, the other languages will follow in due course.

In the following, we give a description of a sample complex term (notice that the German version is a simple term):

Language	Complex Term
English	balance of payments b.o.p. external accounts balance of trade and transfers balance of international payments international balance of payments
German	Zahlungsbilanz
French	balance des paiements
Italian	bilancia dei pagamenti
Spanish	balanza de pagos

4. The Generation Phase

The second Phrase Manager stage in its use for bank terminology implementation is the compilation and generation phase. After having defined the whole range of features and data to be exploited during the analysis approach, the system must first check data consistency, then compile and generate information in form of finite-state tools, as shown in Fig. 3.

The finite-state tools are in fact a collaboration of different elements, one for the morphological analysis, one for the periphrastic detection and one for the complex term word sequence detection. This combination of tools is used to efficiently extract terms from a text.

5. Analysis and Extraction

The text analysis and extraction of terms represent the final step and the interaction with the end-user. The analyzer acts as external interface for the access to the terminology database. Using the collection of finite-state tools generated by the modeling system, it analyzes on-line each text sentence and, in case of FST retrieval, it can deliver as answer whatever information it can get from the database for the specific recognized term. This means that the analyzer must have an access to the original terminology database, in order to prepare a complete answer for the end-user.

It has been explained that terms can appear in different forms, depending on the characteristics associated during the definition.

The approach followed to recognize terms in a text is pattern matching based on regular expressions.

The analysis is divided in different levels:

¥ There is a morphological analysis, which first allows the detection of single parts of a complex term. Here we use an inflection transducer, generated from our morphological module, able to analyze each single word and deliver the corresponding citation form. The more complete transducer at the moment is the German one, which counts about 120,000 lexemes (more than 1 million word forms, which need 1.4 MB of disk space and about 6 MB of memory during run-time).

¥ There is also a periphrastic analysis, which looks for cluster elements, like separable German verbs or, more in general, multi-word verb paradigm forms. Finally there is a word sequence analysis, implemented in a further transducer which contains information about canonical form, all possible word sequences, modifications and transformations of a complex term.

5.1. Examples

The easiest complex terms that can be extracted are of course the ones that do not have any discontinuity, like the former balance of payments or Protected Index Participation Unit (the same in every language), etc.

More complex are the examples where inflection and discontinuity can appear, like "interest rates rise", that will be considered through the following analysis.

The output of the analysis is just a first result that is used to access the database and get more useful information on the specific term. The indexes that precede the canonical form represents the positions (starting by zero) of the single elements within the input text.

Input: yesterday the interest rates rose
(a)
Output: 2 3 4 interest rates rise

Input: because the interest rates have risen
(b)
Output: 2 3 4 5 interest rates rise

Input: how much did the interest rates rise?
(c)
Output: 2 4 5 6 interest rates rise

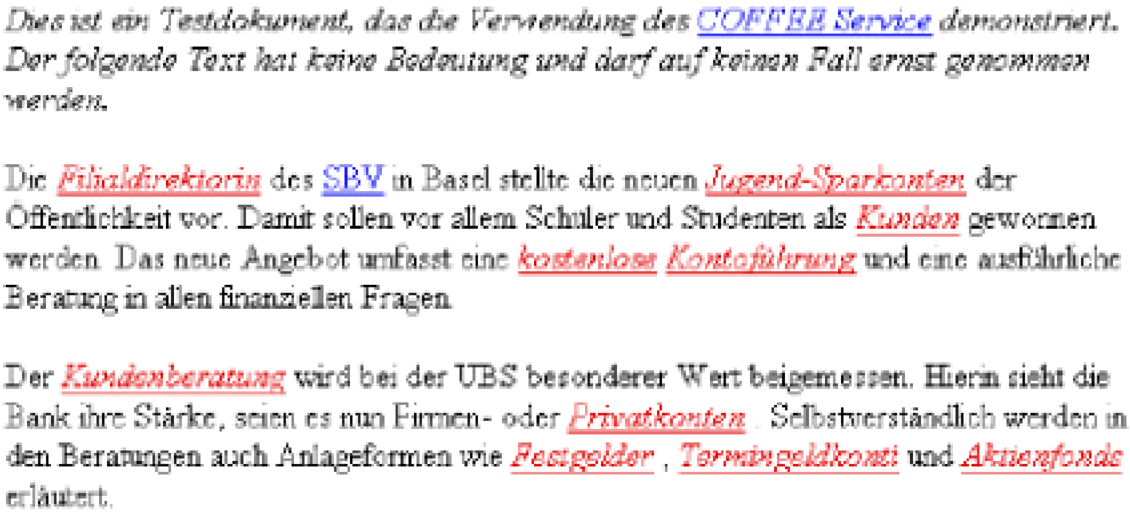
The example (a) contains a term that deviates from its original canonical form ("the interest rates rise") by the fact that some of its elements use different paradigm word forms. This is a morphological deviation and can be detected by a usual morphological transducer.

The example (b) presents a more interesting feature. The verb uses a composite word form ("have risen"). The composite word form is first recognized and used to build a periphrastic cluster, which is then considered as unique graphic word form for the term analysis and extraction.

The example (c) shows that the term is recognized even if its elements are not consecutive. The periphrastic form did rise is in fact extracted and recognized, in order to match the canonical form of the term.

5.2. Intranet Environment

In this paragraph some images of the intranet interface are presented. The first one (Fig. 3) shows an HTML page where some terms have been recognized.



Dies ist ein Testdokument, das die Verwendung des COFFEE Service demonstriert. Der folgende Text hat keine Bedeutung und darf auf keinen Fall ernst genommen werden.

Die *Filialdirektorin* des SBV in Basel stellte die neuen *Jugend-Sparkonten* der Öffentlichkeit vor. Damit sollen vor allem Schüler und Studenten als *Kunden* gewonnen werden. Das neue Angebot umfasst eine *kostenlose Kontoführung* und eine ausführliche Beratung in allen finanziellen Fragen.

Der *Kundenberatung* wird bei der UBS besonderer Wert beigemessen. Hierin sieht die Bank ihre Stärke, seien es nun Firmen- oder *Privatkonten* . Selbstverständlich werden in den Beratungen auch Anlageformen wie *Festgelder* , *Termingeldkonti* und *Aktienfonds* erläutert.

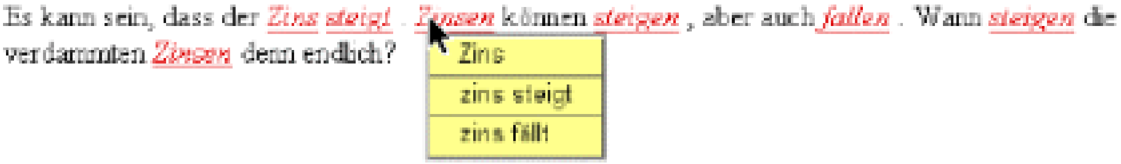
Fig. 3: Recognized Terms

To each recognized term a link to the terminology database has been associated. Before having to access the terminological database to get more information, you can obtain some first quick information simply moving the mouse over the link.

This movement will activate a tooltip, which can deliver:

¥ the citation form, in case of single terms and abbreviations,

¥ or the canonical form in case of complex terms (as shown in Fig. 4).



Es kann sein, dass der *Zins steigt* . *Zinsen* können *steigen* , aber auch *fallen* . Wann *steigen* die verdammten *Zinsen* denn endlich?

Zins

zins steigt

zins fällt

Fig.4 : Tooltip delivering citation form and canonical forms

Notice that in this last example one single word (Zinsen) has been associated with three different terms: the simple term Zins (interest rate) and the two complex terms Zins steigt and Zins f llt . This shows that discontinuity can be applied at several levels, even when one element, as in the case of Zins in Zins f llt , is only implicit.

A further step is to directly access the information in the database, clicking the mouse on the corresponding link. This will deliver information on meaning, source, different translations, context in which the term is used, and other specific observations.

6. Conclusions

The complete data for the analyzer exists only for a German analysis and, partly, for an English one. It will be produced for more languages in the near future, whereas the database information already exists in five languages. The next steps will be the completion of the Phrase Manager specifications for English, and the new specifications for French, Italian and Spanish.

The single finite-state elements are implemented using a generic object-oriented framework. The FST technology has proved to be suited and efficient for such kind of analysis. Moreover, due to its modular implementation, it can be easily integrated into different kinds of services, such as spelling and style checker, or other more advanced NLP tools. Some demonstrations can be viewed at the site http://www.canoo.com

Acknowledgements

We would like to thank the whole Canoo team for its collaboration and advices, particularly Stephan Bopp, Marc Domenig, Urs Hengartner, Dieter Holz, Dirk Krampe and Alexio Moreno.

Bibliographical References

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

Daille B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, In: The Balancing Act:

Combining Symbolic and Statistical Approaches to Language, Judith L. Klavans and Philip Resnik", MIT Press", Cambridge, MA, 1996..

Frantzi K.T., Ananiadou S.: *Extracting Nested Collocations*, In Proceedings of the 16th international conference on computational linguistics, Coling 96, 41-46, 1996.

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

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

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

Pedrazzini S.: Treating Terms As Idioms, In Proceedings of the Sixth International Symposium on Communication and Applied Linguistics, Santiago de Cuba, Editorial Oriente, Santiago de Cuba 1999.

Schenk A.: The formation of idiomatic structures, Proceedings of the First Tilburg Workshop on Idioms, 1989.

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

Verstraten L.: Idioms in Dutch dictionaries, Proceedings of the First Tilburg Workshop on Idioms, 1989

Volk M.: The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems, In: Nico Weber (ed.): Machine Translation: Theory, Applications, and Evaluation. An assessment of the state of the art. St. Augustin: gardez-Verlag. 1998.