Catalina Jiménez / Claudia Seibel
Granada

From expert knowledge representation to controlled language: Flexible definitions for coastal process concepts

1 Introduction

This article is a description of the preliminary results of the research presently being carried out within the framework of the R&D project *Coastal engineering: specialized knowledge representation and generation of terminological resources*[1], funded by the Ministry of Education in Spain.

As its name implies, one of the goals of this research project is the generation of terminological resources for specialized knowledge concepts in the form of trilingual terminological glossaries and dictionaries as well as multimedia databases for the specialized domain of Coastal Engineering. These resources are based on the representation of the conceptual structure of this domain, relevant concepts, defining characteristics, and relations.

The preliminary conceptual design is derived from terminographic hierarchies elaborated on the basis of conceptual data extracted from a corpus of engineering texts in English, Spanish, and German as well as specialized dictionary entries. The conceptual macro-representation of the field takes the form of a prototypical cognitive event frame that relates the various conceptual roles by means of semantic relations (Faber 2003). This event consists of a set of interrelated categories, the names of which can be regarded as the generic terms defining the other members of the category. The interrelations between concepts are made explicit in their semantic definitions.

Defining a concept and the terms associated with it signifies making a linguistic description that reflects the information that experts possess and transmit when the concept is activated in a text. This also includes the speaker's communicative intention (e.g. *describe, explain, contract, criticize,* etc.), his/her perspective (e.g. areas and sub-areas implicitly referred to), and the specification of the difference between the concept being defined and others belonging to the same domain. As a result, a definition is a kind of knowledge representation, and in this case, the representation is of expert knowledge related to Coastal Engineering.

One of the most crucial factors here is a definitional metalanguage for each category and the concepts belonging to it. One advantage of basing such a metalanguage on natural language is that when it is sufficiently constrained, it becomes highly controlled. In this article we propose the creation of a controlled language that can be used as a language for the conceptual representation of

[1] This research bas been carried out within the framework of the R&D project "Ingeniería de puertos y costas: Estructuración de Conocimiento y Generación de Recursos Terminológicos" (ref. no. BFF2003-04720) funded by the Spanish Ministry of Education.

concepts in a specialized knowledge base, which would take into account possible textual activations of each term associated with a concept.

In this article we describe the methodology that we have used for the specification of a variable definition, depending on the knowledge, expectations, and needs of the potential user. This methodology is applied to the subdomain of coastal processes.

2 Terminology management based on corpus analysis

One of the greatest challenges in Terminology is the establishment of theoretical premises that explain and describe new forms of knowledge representation. In this sense, Schmidt-Wigger (1999:1) states that terminology constitutes the major part of a technical document.

In fact, an important advance in the discipline of Terminology in the past decade has been the configuration of a model of lexical analysis, the results of which can be applied to the creation of a definitional prototype that is psychologically adequate and in accordance with psycholinguistic models of information processing (Faber/Mairal 1999; Faber 2002; Seibel/Jiménez 2004).

However, as a general rule, little is said regarding a viable methodology that can be used to elaborate definitions based on the extraction of conceptual, terminological, and pragmatic information from a corpus compiled with a specific objective in mind[2].

With the help of information extracted from specialized dictionaries as well as a corpus of specialized texts, we have created a highly controlled definitional language, capable of being modelled with a view to different types of possible users. The reason for this flexibility and possibility of adaptation lies in the inclusion of pragmatic information, which is always related to the text type in which it will be used, the text type from which it has been extracted, as well as the senders and receivers of these texts.

During the compilation of the corpus special emphasis was made on the variation and statistical distribution of different text types. These are in consonance with the social context of specialized communication in the area under study. In this sense it was very important to control and label each text so as to be aware of its level of specialization and abstraction as well as its prototypical receiver.

The language used in the elaboration of each type of definition is based on that found in the corpus, and thus, is used by experts who address potential groups of receivers. The use of real language provides us with the tools and means to adapt terminological definitions to specific groups of text users, and consequently, to create a *variable definition.* The following section discusses the concept of controlled languages within the context of recent research in this area.

---

[2] An exception to this rule is Pérez Hernández (2004).

3 Knowledge representation in controlled language: an overview

The notion of knowledge representation is actually a relatively simple one. It has to do with writing down in some language or communications medium descriptions or pictures that correspond to the world or a state of the world. It goes without saying that the language used should be highly constrained, and its semantic and syntactic components should have descriptive adequacy. Within the domain of Artificial Intelligence, a coherent description is regarded as a set of recurrent elements, which an intelligent machine can use and define. This is the basic premise that relates knowledge representation and controlled language.

According to Mahesh and Nirenburg (1995) the primary components of a good knowledge representation are the following:

- a language of representation
- inferential capacity
- domain knowledge

These authors generally affirm that such a representation should have expressive adequacy and above all, be able to reason efficiently. In consonance with these assertions, knowledge representation experts seem to agree that the criteria for evaluating the adequacy of a representational language are the following:

- logical capability to enhance the expression of knowledge
- heuristic potential or the ability to solve problems (inference)
- notational simplicity with a view to facilitating comprehension

It is evident that these criteria are necessary for the objectives proposed in our study, one of which is to create a specialized knowledge base. However, nothing is explicitly said about how to go about elaborating a coherent, controlled language that would fulfil these conditions. By *coherent* we mean that such a language should be based on how experts express themselves when they elaborate texts for different groups of receptors with varying levels of specialized knowledge.

In line with this, translators and terminologists that design controlled languages for computer applications and translation tools affirm that characteristics of controlled language are monosemy, consistence, comprehensibility, and precision. For example, if the description of a new product were written in such a language, the text could be easily translated without any problems of cultural understanding. Such a translation could presumably be carried out almost automatically, something that would evidently save money and be very cost-effective.

The issue here is to discover if the use of controlled language truly facilitates comprehension; if its use is an advantage for those who must write texts in a language that is not their mother tongue; and if controlled language avoids ambiguity. Shubert et al. (1995:360ff.) carried out an experiment, the results of

which showed that non-native text receivers understand controlled-language texts better. At the terminological level, this was due to the following factors:

- the use of terms with a higher frequency
- the use of shorter, less technical terms
- the use of internationalisms or words that have cross-linguistic similarity
- the absence of synonyms

The syntactic level had the following characteristics:

- short, concise sentences
- the codification of one expression per sentence
- the use of the imperative to give instructions
- the use of the active voice to describe actions

Despite the advantages of controlled language in the elaboration of specialized texts, even those who endeavour to use it are often unable to follow the above rules.

Largely due to his experience in European terminology projects, Schmidt-Wigger (1999:3) affirms that terminological coherence depends on the terms being well defined and systematically interrelated. She affirms that terminology control always depends on a web of defined terminology, something that was neither available from the industrial partner, nor could be defined by the linguistic partner of the project.

In contrast to others who enthusiastically endorse controlled language, Janowski (1998:1) is one of the few who mentions its risks and possible secondary effects. He states that it is necessary to be aware of the danger inherent in excessive simplification, and the elimination of colloquialisms.

In our opinion, a controlled language is far from being a single, immutable entity. It should rather be regarded as something dynamic, whose variability must always be based on a coherently defined set of terms. Coherence in this case is the product of structural consistence and recursiveness.

4 From knowledge representation to lexical definition: a case study

When working with a corpus or with a semi-automatic textual analysis program[3] it is necessary to isolate the subdomain and be aware of its most representative concepts. In our case, the most relevant concepts describing the subevent were erosion, transport and sedimentation[4].

As can be seen in figure 1, this domain can be categorized in terms of the following macro-roles: an AGENT *Meer,* by means of INSTRUMENTS such as *Wellen*, *Strömungen* and *Gezeiten*, that generates the PROCESSES of *erosion*, *transport*

[3] In our case WordSmith Tools (http://www.oup.co.uk).

[4] See figure 1 which represents the subdomaine of the *marine processes*.

and *sedimentation,* which act on the coastline conceptualized as a PATIENT. Each process produces effects on the coastline, which are perceived as RESULTS.

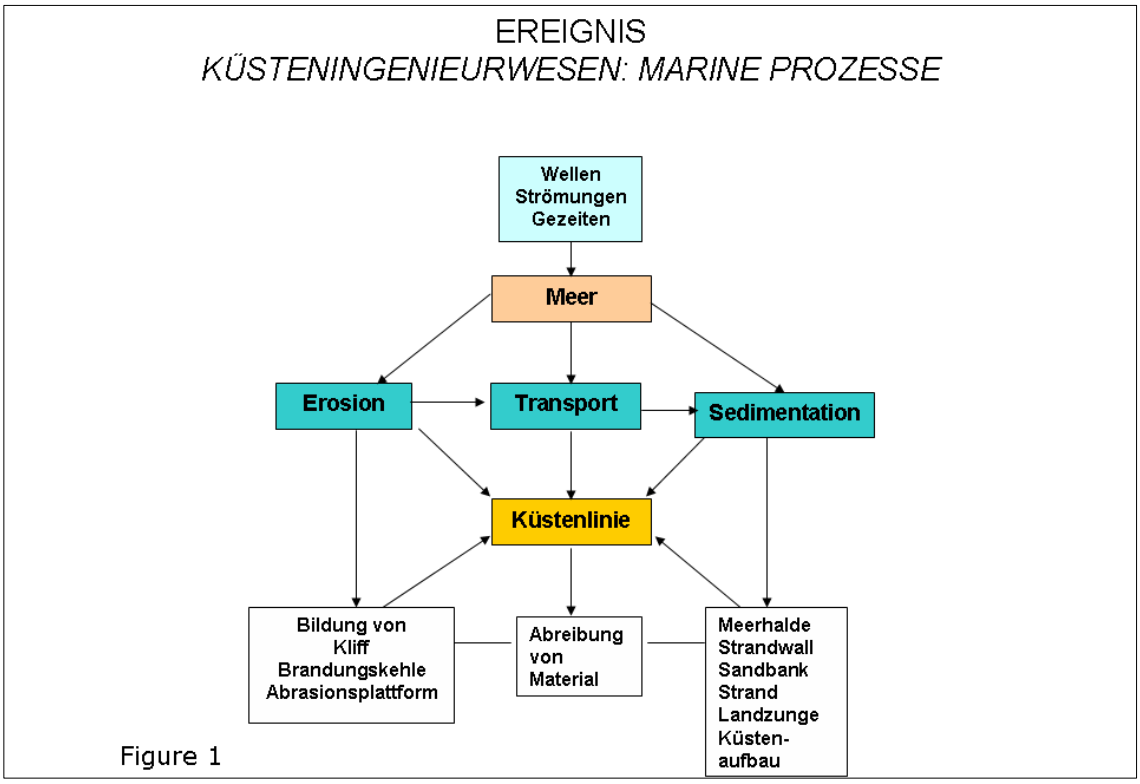

Figure 1

Once the subdomain is thus defined, concordances are extracted for the terms. Figure 2 shows the concordances for the concept of *erosion,* from which it is possible to extract the following types of information:

- Conceptual information
- Terminological information
- Syntactic information
- Pragmatic information

An example of conceptual information can be found below:

> Durch die Aufspülung wird das durch Erosion abgetragene Material teilweise [...] (1)
>
> [...] an Stellen intensiver Erosion (Abtrag stärker verwitterten Materials) [...] (4)
>
> [...] an der Basis des Kliffs. Bei weiterer Erosion bricht das Gestein darüber nach [...] (15)

These concordances provide information regarding the semantic relations between different definitional components. When they explain the causes and effects of the process of erosion, they help the terminologist to construct a dynamic definition, which includes all necessary data for the activation of the mental image of erosion, as well as the data necessary for the description of expert knowledge.

A second example of information that can be directly included in the definition can be found in the concordances below, which offer a description for the concept of *Kliff* and its terminological hyponym *aktives Kliff*:

Kliff: Durch Erosion entstandenes Steilufer. (11)

„aktives Kliff", bei dem die Erosion andauert [...] (12)

Kliffs sind durch Erosion entstandene Landschaftsformen (13)

Konkordanzen zu EROSION,
erstellt anhand des Korpusanalyseprogramms WordSmith Tools

WordSmith Tools	
N	Concordance
1	n. Durch die Aufspülung wird das durch Erosion abgetragene Material teilweise ersetz
2	legenen Ort statt. Deswegen wird durch Erosion Material von oben nach unten verlag
3	ne höhere Fließgeschwindigkeit erfolgte Erosion auf dem zuerst sedimentierten Mater
4	Entwicklungszeit) an Stellen intensiver Erosion (Abtrag stärker verwitterten Material
5	te und den Rändern. Dabei findet keine Erosion statt, da kein Material bewegt wird. I
6	ende Materialversorgung der unter Lee-Erosion leidenden angrenzenden Strand-Abs
7	nsport von Material (transportlimitierte Erosion) zur Ausbildung einer Verwitterung
8	h zu einer ständigen Akkumulation und Erosion von Material. Dadurch ist die Gestein
9	l stark belastet. Bodenbearbeitung und Erosion führten später zur Nivellierung der Au
10	bildet und in Folge von Ablagerung und Erosion geprägt, wobei sowohl die Materialzu
11	egend rolligem Material. Kliff cliff Durch Erosion entstandenes Steilufer. ANMERKUNG
12	nterscheidet „aktives Kliff", bei dem die Erosion andauert und das dadurch landwärts
13	lle: Kelletat, D. (1999) Kliffs sind durch Erosion entstandene Landschaftsformen, mit
14	rm, auch Felsschorre genannt. Erneute Erosion am Kliff bewirkt neuen Schutt und da
15	hle an der Basis des Kliffs. Bei weiterer Erosion bricht das Gestein darüber nach, und

Figure 2

Moreover, the terminological information in the concordances (7), (8), (10) does not need specification since it is evident that it relates terms that are more or less directly associated with *erosion*: transportlimitierte *Erosion* (a hyponym) and other cognitively related terms, such as *Akkumulation* and *Ablagerung*.

The syntactic information is present in practically all of the concordances since it illustrates how experts use the term *erosion* and other cognitively-related ones. There is what could be considered an excessive use of the preposition *durch* in combination with *Erosion.* This points to the activation of the thematic role of INSTRUMENT in most of the syntactic constructions. The frequent use of verbs such as *bewirken* or *durch … geprägt* also indicates tendencies in syntactic collocations. (See section 4 for the analysis and implications of pragmatic information in concordances.)

The following examples not only show examples of conceptual, terminological, collocational, syntactic and pragmatic information, but also describe the relation between *Erosion* and *Abrasion,* its most direct hyponym:

An den Steilküsten wirkt die Abrasion, also die marine Erosion oder der Abrieb, durch die Arbeit der Brandungswellen an der Basis eines Kliffs.

Die hydraulische Kraft des Wassers prägt die Küsten durch verschiedene Vorgänge.

Abrasion ist die abtragende Wirkung der Prozesse der Brandung an der Küste. Sie wirkt nicht nur im Locker-, sondern auch im Festgestein.

Insbesondere das Kollidieren von Frachtmaterial mit Hindernissen und das damit verbundene Abschaben trägt zur Erosion bei (Abrasion).

Infolge von Abrasion und Wellenwirkung entsteht allmählich eine parallel zur Steilküste verlaufende Brandungskehle an der Basis des Kliffs.

Durch die Erosion wird das Gestein am Fuß abgetragen, es bildet sich eine Hohlkehle.

Bei der durch die Brandung entstehende Erosion des Untergrundes kann es zur Ausbildung von Abrasionsflächen oder Abrasionsplatten kommen.

After the initial extraction of different types of information, the next step is to use this data to formulate a schema or template valid for all the concepts belonging to the specialized area being described. The concordances analyzed in our corpus give us the following template:

X is a TYPE-OF coastal PROCESS that has a CAUSE, affects a PATIENT, and produces a RESULT.

This schema is the natural-language abstraction of the linguistic and cognitive elements of the definition. As a result, these elements are activated by the text senders, and are regarded as those that should be evoked in the receivers. If we return to figure 1, we can see that the patient role is practically always taken by the Küstenlinie, while the results produced are invariably types of Kliff, Sandbank, Brandungskehle, etc.

Figure 3 presents the conclusions regarding the meaning of the concepts of Erosion and Abrasion. Thanks to their shared cognitive structure, it is evident that both concepts conform to the same template. At the same time, it is made evident that ABRASION is a concept that is immediately subordinate to EROSION, in that it inherits its definitional information and structure, while introducing greater conceptual specification, as can be seen in the underlined words in the following examples:

EROSION: (process) bestehend in der Abreibung von Material

ABRASION: (process) bestehend in <u>flächenhafter Abreibung</u> von Material

EROSION: (patient) beeinflusst die Küste

ABRASION: (patient) beeinflusst die Steilküste

However, despite having elaborated natural-language definitions based on a corpus of authentic texts, with a high degree of systematicity, we have not as yet offered a definition that can be adapted to different groups of users.

5 The variable definition: first step towards a controlled language

Seibel (2004) conclusively shows that to create a definition targeted at a specific group of receivers, it is first necessary to formulate definitional hierarchies as described in the previous section. Concordances are thus generated from texts labelled in consonance with potential groups of receivers, and this data is used to provide information regarding different text types. Consequently, the definitional hierarchies derived in this way from corpus structure, reflect the social context of the texts produced in the knowledge domain of Coastal Engineering.

Figure 4 shows concordances which in this case reflect the relation PATIENT-PROCESS, activated in the concept of *sedimentation*. We have observed that this relation is present in all of the texts compiled on *maritime processes.* As expected, in semi-specialized texts as well as those for popular consumption, we found that the language used was very clear, descriptive, and concrete, whereas the language used in technical texts was very abstract. When the concordances are organized in hierarchies in terms of text type (see figure 4), it is possible to observe how the language used differs in terms of technicality. The concordances from texts targeted at receivers without any specialized knowledge are the following:

> Es bilden sich Flachküsten mit breiten Sand- oder Kiesstränden. [...]
>
> An Flachküsten, an denen sich das Land nur langsam zum Meer hin absenkt [...]

These present a vivid contrast to highly technical texts in which we find linguistic elements such as those found in the concordances below:

> Die Küste kann in Abschnitte mit Sedimentdefizit (Erosion) bzw. mit Sedimentüberschuss (Akkumulation) unterteilt werden. [...]
>
> [...] tritt in einem capacity-limited system Erosion dort ein, wo die Transportkapazität zunimmt,[...]

In the above concordances it is also possible to perceive the codification of the explicative speech act. In specialized texts this can be seen in the explanations and clarifications in parenthesis, whereas in the extremely specialized ones terms are given in English without any explanation (in einem capacity-limited system), and there is the evident assumption that the text receivers will understand the conceptual content of the text.

This is precisely the type of conceptual, terminological, and syntactic information that will be used in the elaboration of definitions in our knowledge base. In consonance with this, one type of definition or another will activated, depending on the expectations and needs of the user. Since these are maritime processes, these concepts presumably share the same type of definitional schema. This schema contains the basic information template that includes the information required to understand the concept, and so the only thing that will vary is the type of language used to fill in the slots, and which will vary depending on the receiver.

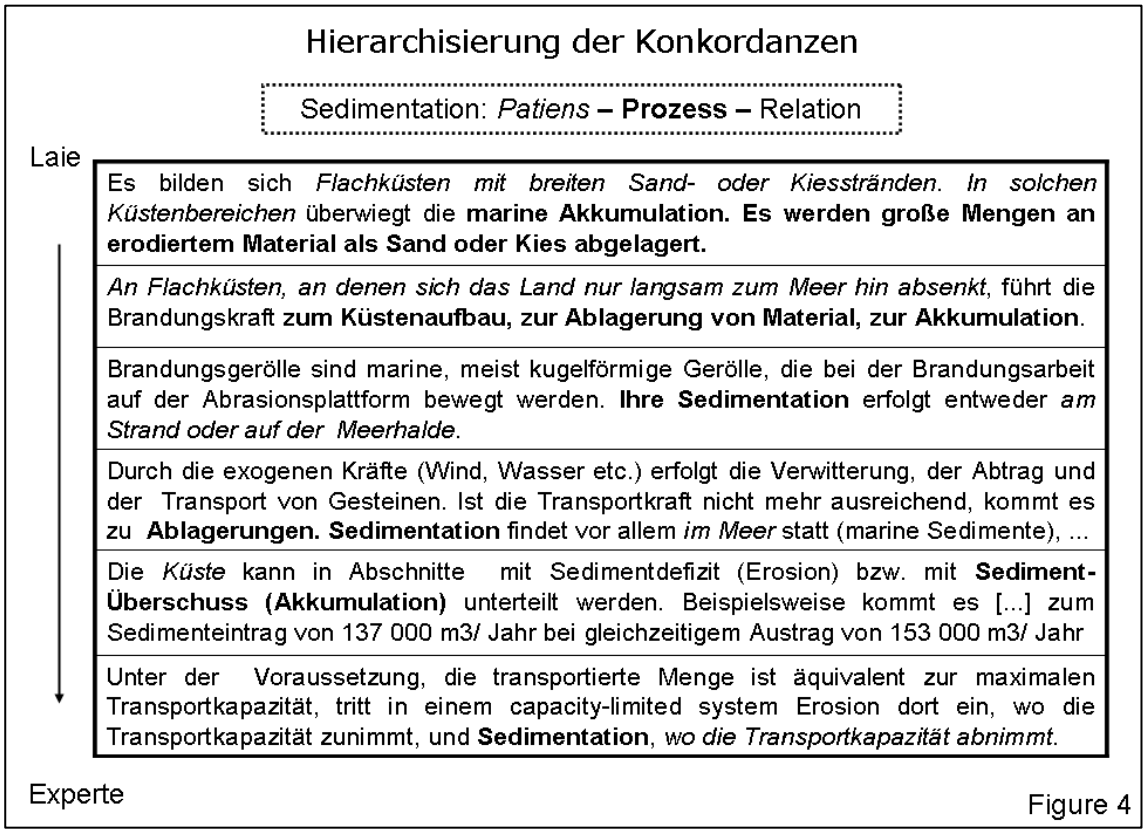

Figure 4

As a result, figure 5 shows different types of linguistic designations for the same concept in the same definitional format. For example, in the conceptual relation TYPE-OF, the expert definition uses the term *mariner Prozess* whereas the definition for the non-specialist would be *Meeresprozess*. The terms used for the macro-role PROCESS also show meaningful differences:

- EXPERTE: ... bestehend in der Ablagerung und Akkumulation von Material nach dessen Erosion und Transport
- LAIE: ... vorher abgetragenen Sand oder Kies ablagern und ansammeln

Finally, figure 6 is a contrastive representation of the hierarchization that appears in the different text types. This type of controlled language regards the concept of erosion. The order is according to degree of difficulty and abstraction. The term appears in collocations with the relations that are activated in the structure of the definition (process, patient, instrument, agent, and result), combined with the different linguistic structures that could be used by a non-expert in the field (e.g. a journalist, translator, or reviewer) if he/she wished to activate them in order to construct a text in this language for a specific group of receivers.

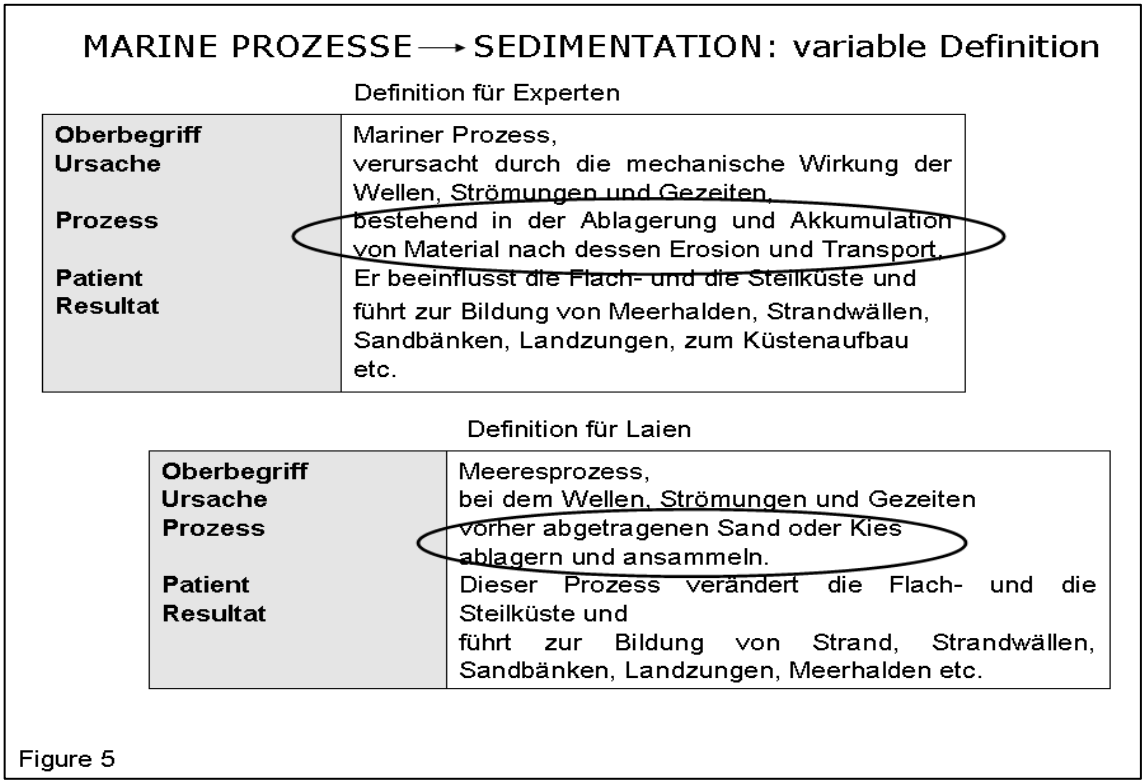

Figure 5

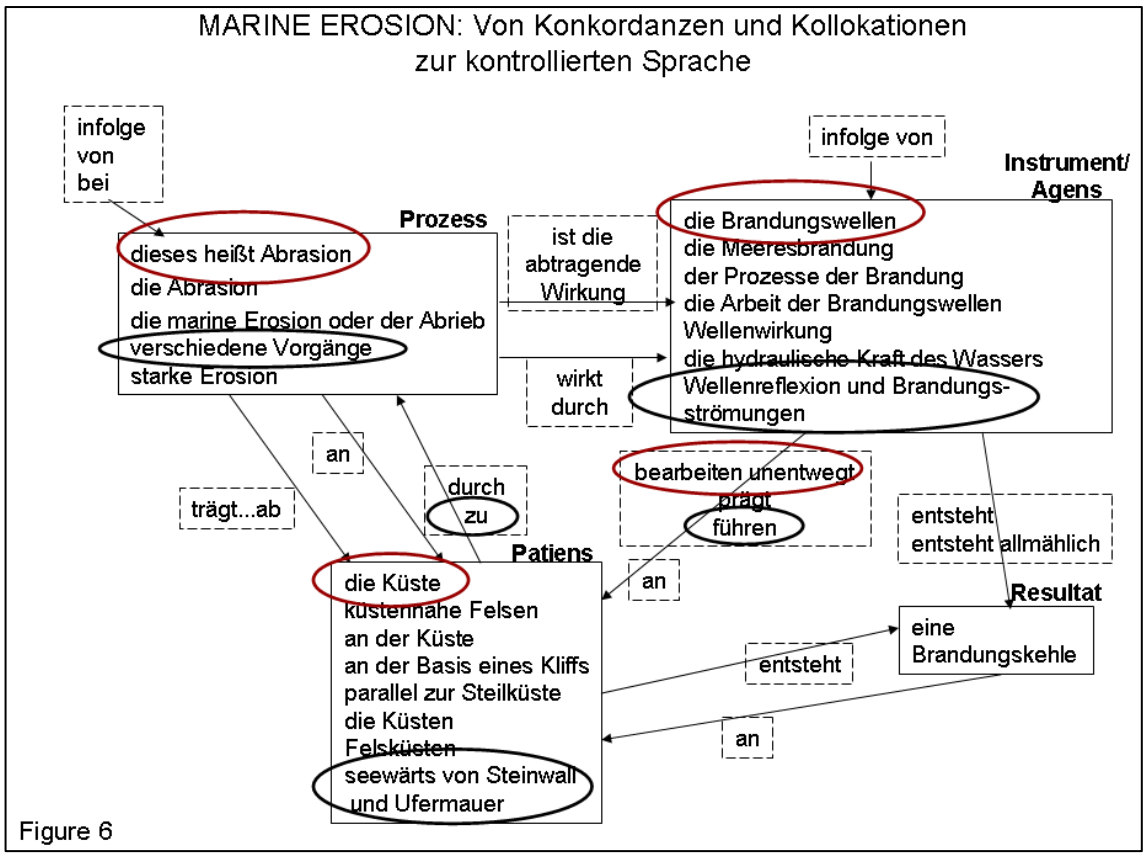

Figure 6

In reality, this could be called a stylistic micro-grammar for the terms designating the concept of COASTAL EROSION. If a non-expert followed the arrows and

selected the parts of the proposed itinerary, he/she could presumably construct the following text with the elements given in the upper part of figure 6:

> Die Brandungswellen bearbeiten unentwegt die Küste. Dieses heißt Abrasion.

In the case of a text for experts, the elements selected would naturally vary. The linguistic constructions used would be those found in the lower part of figure 6:

> Wellenreflexion und Brandungsströmungen führen seewärts von Steinwall und Ufermauer zu verschiedenen Vorgängen.

6 Conclusions

In this paper, we have shown how controlled language can be extracted from a corpus of authentic texts that have previously been labelled according to their degree of abstraction, and thus categorized in terms of groups of potential receivers. This type of controlled language is hierarchically configured according to these previously stated parameters, and can be used to construct a meaning definition for each concept, which will vary according to the knowledge and expectations of the user.

Each definition is elaborated according to a template containing both syntactic and semantic information. The template will vary, depending on the conceptual category. On this basis, the following pragmatic factors must be taken into account in the construction of a variable definition:

- The characterization of the text-type in which the concept is activated;
- The cognitive distance between the "neutral" term and the one being defined, as well as the reasons justifying this distance;
- The inclusion of a pragmatically adequate language which reflects both new and shared information for the text receivers.

References

Allen, Jeff (1999): "Different Types of Controlled Languages." TC-Forum. http://www.tc-forum.org/topiccl/cl15diff.htm.

Faber, Pamela (2000): *La aplicación del MLF a la terminología*. Postgraduate seminal given at the University Pompeu Fabra, Barcelona, Spain (1998-2000).

Faber, Pamela (2002): "Oncoterm: sistema bilingüe de información y recursos oncológicos." Alcina Caudet, Amparo / Gamero Pérez, Silvia (Hrsg.): *La traducción científico-técnica y la terminología en la sociedad de la información*. Castelló de la Plana, Publicacions de la Universitat Jaume I, 177-188.

Göpferich, Susanne (1998): *Interkulturelles "Technical Writing". Fachliches adressatengerecht vermitteln. Ein Lehr- und Arbeitsbuch*. Tübingen: Narr.

Göpferich, Susanne (2002): *Textproduktion im Zeitalter der Globalisierung. Entwicklung einer Didaktik des Wissenstransfers*. Tübingen: Stauffenburg.

Janowski, Wladyslaw (1998): *Controlled Language – Risks and Side Effects*. TC-Forum. http://www.tc-forum.org/topiccl/cl14cont.htm (24.04.2007).

Jiménez Hurtado, Catalina (2001): *Léxico y Pragmática*, Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, Bd. 5. Frankfurt am Main: Peter Lang.

Jiménez Hurtado, Catalina / Seibel, Claudia (2004): "El lenguaje controlado para una definición variable: 'no ambiguity through homonyms, no redundancy through synonyms'." Faber, Pamela / Jiménez Hurtado, Catalina / Wotjak, Pert (Hrsg.): *Léxico especializado y comunicación interlingüística*. Granada: Granada Lingvistica, 117-130.

López Rodríguez, Clara Inés (2001): *Tipología textual y cohesión en la traducción biomédica inglés-español: un estudio de corpus*. Granada: Editorial Universidad de Granada.

Mahesh, Kavi / Nirenburg, Sergei (1995): "Semantic classification for practical natural language processing." Proc. Sixth ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting. Chicago: IL.

Moreno Ortiz, Antonio (2000): "Managing conceptual and terminological information in a user-friendly environment". *Proceedings of OntoLex 2000. Workshop on Ontologies and Lexical Knowledge Bases*, September 2000. Sophia: Bulgaria.

Mügge, Uwe (2002): "Möglichkeiten für das Realisieren einer einfachen Kontrollierten Sprache", *Lebende Sprachen* 3, 110-114.

Pérez Hernández, Chantal (2000): *Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas*, unpublished PhD thesis, University of Malaga.

Reuther, Ursula (1998): "Controlling Language in an Industrial Application." *Proceedings of CLAW'98*. http://www.iai.uni-sb.de/docs/clrev.pdf (24.04.2007)

Schmidt-Wigger, Antje (1999): "Term Checking through Term Variation." *Proceedings of TKE '99*, 23-27.8.99, Innsbruck. http://www.iai.uni-sb.de/docs/tke.pdf (24.04.2007)

Seibel, Claudia (2004): *La codificación de la información pragmática en la estructura de la definición terminológica*. Granada: Editorial Universidad de Granada.

Shubet, Serene / Spyridakis, Jan / Holmback, Heather / Coney, M. B. (1995): "The Comprehensibility of Simplified English in Procedures." *Journal of Technical Writing and Communication*, 25, no. 4, 347-36.

Temmerman, Rita (2000): *Towards new ways of terminology description: the sociocognitive approach*. Amsterdam/Philadelphia: John Benjamins.

Temmerman, Rita (2001): "Sociocognitive terminology theory." Cabré, M. T. / Feliu, J. (Hrsg.): *Terminología y cognición. II Simposio Internacional de Verano de Terminología*. Barcelona: IULA, 75-92.

Wittwer, Michael (2001): "Fachsprachliche Besonderheiten der verschiedenen popularisierenden Fachtextsorten in der Pädiatrie." *Fachsprache. International Journal of LSP*, 23. Jahrgang, Heft 1-2, 71-91.