Workshop Programm

14:30 – 14:45	Rute Costa, Raquel Silva, Maria Teresa Lino	ExtracTerm: an extractor for portuguese language
14:45 – 14:55	Debate	
14:55 – 15:10	Masaki Murata, Qing Ma, Tamotsu Shirado, Hitoshi Isahara	Database for Evaluating Extracted Terms and Tool for Visualizing the Terms
15:10 – 15:20	Debate	
15:20 – 15:35	Michael Carl, Maryline Hernandez, Susanne Preuβ, Chantal Enguehard	English Terminology in CLAT
15:35 – 15:50	Debate	
15:50 – 16:05	Pierre Auger	Terminotics and Research in Canada and Quebec: an actual overview
Coffee Break		
16:30 – 16:45	Lotte Weilgaard	Valency Patterns of Danish Verbs as Terminological Knowledge Patterns
16:45 – 16:55	Debate	
16:55 – 17:10	Elisabeth Marsham	The Cause-Effect Relation in a French-Language Biopharmaceuticals Corpus: Some Lexical Knowledge Patterns
17:10 – 17:20	Debate	
17:20 – 17:35	Lone Bo Sisseck	Semantic Relations between Concepts in Danish Domain Specific Texts
17:35 – 17:45	Debate	
17:45 – 18:00	Judit Feliu, John Jairo Giraldo, Vanesa Vidal, Jorge Vivaldi, M. Teresa Cabré	*The GENOMA-KB project: a concept based term enlargement system*
18:00 – 18:10	Debate	
Coffee Break		
18:30 – 18:45	Bodil Nistrup Madsen, Hanne Erdmann Thomsen, Carl Vikner	CAOS – a support system for ontology structuring
18:45 – 18:55	Debate	
18:55 – 19:10	Magnar Brekke	*KB-N: Computerized extraction, representation and dissimination of special terminology*
19:10 – 19:20	Debate	

Workshop Organizers

Rute Costa - Universidade Nova de Lisboa – CLUNL - Portugal
Lotte Weilgaard - University of Southern Denmark - Denmark
Raquel Silva - Universidade Nova de Lisboa – CLUNL - Portugal
Pierre Auger - Université Laval - Canada

Workshop Program Committee

Bertha Toft - University of Southern Denmark – Denmark
Jacques Ladouceur – Université Laval – Canada
Maria Teresa Lino – Universidade Nova de Lisboa - Portugal
Marie-Claude l'Homme – Université de Montréal – Canada
Philippe Thoiron – Université Lumière 2 – France
Rosa Estopa – Universidade de Pompeu Fabra – Spain

Table of Contents

Author Index

Auger, Pierre
Brekke, Magnar
Cabré, Maria Teresa
Carl, Michael
Costa, Rute
Enguehard, Chantal
Feliu, Judit
Giraldo, John Jairo
Hernandez, Maryline
Isahara, Hitoshi
Lino, Maria Teresa
Ma, Qing
Madsen, Bodil Nistrup
Marsham, Elisabteh
Masaki Murata,
Preuβ, Susanne
Shirado, Tamotsu
Silva, Raquel,
Sisseck, Lone Bo
Thomsen, Hanne Erdmann
Vidal, Vanesa
Vikner, Carl
Vivaldi, Jorge
Weilgaard, Lotte

ExtracTerm: an extractor for Portuguese language

Rute Costa, Raquel Silva, Maria Teresa Lino
Universidade Nova de Lisboa - CLUNL
Unit Research – Lexicology, Lexicography and Terminology
Avenida de Berna 26 – C
1069 – 061 Lisboa – Portugal
m.rutecosta@mail.telepac.pt; raq.silva@fcsh.unl.pt; unl.tlino@mail.telepac.pt

Abstract

We intend to describe the tagger **EtiqueLex** and **ExtracTerm**, an extractor for Portuguese language. The results, that we consider rather satisfactory, obtained from the running of these two products are, in part, a direct consequence of the establishing of rigorous criteria as far as selecting the texts is concerned. On the other hand, the identification, from a theoretical point of view, of the linguistic entities upon which our study fell, multlexemic terminological units, has allowed us to identify the most frequent terminogenic structures in our *corpus*.

ExtracTerm is a dynamic software. The dictionaries at the base of its performance, that is, the dictionary of inflected forms and the dictionary of locutions are open dictionaries, to which it is possible to add information at any moment, every time it is proven necessary.

Introdução

The basis for the ellaboration of two software programs on language for special purposes automatic treatment, a tagger – EtiqueLex – and an extractor – ExtracTerm – is a set of theoretical and methodological principles of fundamental importance.

The results, that we consider rather satisfactory, obtained from the running of these two products are, in part, a direct consequence of the establishing of rigorous criteria as far as selecting the texts is concerned. On the other hand, the identification, from a theoretical point of view, of the linguistic entities upon which our study fell, multlexemic terminological units, has allowed us to identify the most frequent terminogenic structures in our *corpus*.

From the observation of such *corpus*, we have established recognisance rules that allowed ExtracTerm to recognise all structures previously contemplated on the dictionary of recognisance rules we have conceived.

Also the disambiguation rules were conceived from the observation of the results obtained with ExtracTerm.

ExtracTerm is a dynamic software. The dictionaries at the base of its performance, that is, the dictionary of inflected forms and the dictionary of locutions are open dictionaries, to which it is possible to add information at any moment, every time it is proven necessary.

EtiqueLex: a tagger for portuguese language

However, ExtracTerm can only work on tagged text.

Taking a special purpose-based dictionary as a core set we tagged each of the flexed forms constituting it, bearing in mind that each one can be attributed several tags:

formas	**etiquetas**
colorida	Adj:f:s
colorida	Pp:f:s
coloridas	Adj:f:pl
coloridas	Pp:f:pl
colorido	Adj:m:s
colorido	Pp:m:s
coloridos	Adj:m:pl
coloridos	Pp:m:pl

On the other hand, we have put aside the possibility of attributing tags to the constituents of the grammatical locutions, avoiding the tagger to identify, for instance, the element *recurso* as a [N] in the sequence *com recurso a*. For that reason, we have created a dictionary of locutions containing the following information:

por	oposição a@Loc
por	oposição à@Loc
por	oposição ao@Loc
por	oposição aos@Loc

These are the tagging results:

por_oposição_ a [Loc]
por_oposição_à [Loc]
por_oposição_ao [Loc]
por_oposição_aos [Loc]

Admitting the hypothesis in which EtiqueLex attributes multitaggs to the majority of the flexed forms is, on the other side, to assume the need to create morphosyntactic rules allowing the disambiguation in order to optimise the results obtained from terminogenic matrix previously set out from corpus observation.

EtiqueLex adds to each form every possible grammatical function in different contexts, making it possible that a form be marked with several metalinguistic tags. This procedural phase is similar to the first one of the ExtracTerm, which was conceived to extract information from tagged text. Therefore, EtiqueLex has the particularity of functioning as an autonomous tool or as module of ExtracTerm.

ExtacTerm: Disambiguation rules

After tagging is complete, ExtracTerm departs from the dictionary of typologies, which contains the structures that we are interested in identifying inside the *corpus*. When a form is doubly tagged, ExtracTerm extracts the same sequence twice, if it corresponds to one of the matrixes contained on the typologies dictionary.

This being so, for sequence *abordagem estatística*, in which *estatística* is followed by tag [N:f:s] and tag [Adj:f:s], ExtracTerm applies two recognisance rules: {N:f:s + N:f:s} and {N:f:a + Adj:f:s}, unnecessarily extracting the same sequence twice, for in this context the classification [N:f:s] for the lexeme *estatística* is wrong.

However, this erroneous classification does not affect the result of the extraction, in as much as the specialist is only interested in knowing if the linguistic unit extracted is a denomination or not, independently of the grammatical classification given to each of its constituent elements.

Nevertheless, such a double extraction is useless: it overloads the results, increasing, naturally, the volume of information to be de-codified**.**

If the result is not affected in this context, there are other situations when that is not the case. On the text:

[…] permite [V:cj] a [Art:f:s] [Pron:pess:3p:f:s] [Pron:dem:3p:f:s] [Prep1:a] extracção [N:f:s] dos [Prep2:dos:m:pl] níveis [N:m:pl] de [Prep1:de] reflectância [N:f:s] do [Prep2:do:m:s] coberto [Adj:m:s] [Pp:m:s] vegetal [Adj:2gen:s] [N:m:s]. Contudo [Conj] , os [Art:def:m:pl] [Pron:poss] [Pron:dem] resultados [N:m:pl] obtidos [Pp:m:pl] [Adj:m:pl] se_bem_que [Loc] elucidativos [Adj:m:pl] quanto_aos [Loc] níveis [N:m:pl] de [Prep1:de] actividade [N:f:s] clorofilina [Adj:f:s], […]

we have identified various wrong-tagging situations, which have influenced the extraction negatively

(1) níveis [N:m:pl] de [Prep1:de] reflectância [N:f:s] do [Prep2:do:m:s] coberto [Adj:m:s] [Pp:m:s] vegetal [Adj:2gen:s] [N:m:s].
(2) resultados [N:m:pl] obtidos [Pp:m:pl] [Adj:m:pl]

On the first example, lexeme *coberto* is doubly badly tagged. While building the dictionary of inflected forms, the form *coberto* was not marked with tag [N], because we do not recognise it as lexeme for special purposes out of context**.**

The immediate effect of such wrong tagging was reflected on ExtracTerm's incapacity to extract the multilexémic terminological unit *coberto vegetal*, because the tag automatically given to this sequence does not correspond to any pre-determined matrix.

Also lexeme *vegetal* is doubly tagged, but correctly, although in this context tag [N] is inadequate, once it is unequivocally an [Adj].

In the case of the last example, sequence *informação obtida* is the outcome of the application of an elemental rule {N:m:pl + Adj:m:pl}. Tag [Pp] is ignored, once it is not one of the master constituent classes of multilexemic terminological units withheld, and this way a free sequence of lexemes is obtained. At this stage, ExtracTerm does not withhold the information needed to distinguish a past participle with verbal value and a past participle with adjectival value, what results in a non-productive extraction.

Our aim is to increase to the maximum linguistic correction, in order to decrease the number of mistakes on the extraction. For that effect, it is pivotal to reduce the ambiguity caused by the attribution of multiple tags to a same form, because the more mono-tagged the forms, the more possibilities ExtracTerm has of improving extraction quality.

This time, it is necessary to establish rules that allow the desambiguation between different grammatical classes**.**

To carry out such proceedings, we had to deliberate on which grammatical classes the disambiguation rules must fall upon and when to apply them.

So, we have opted for associating the disambiguation rules to the tagging process. EtiqueLex identifies the lexical contexts of the grammatical classes to be disambiguated and applies the rules, in order to give a single tag to each form in a given context.

From the moment the *corpus* is partially disambiguated, ExtracTerm has fewer tags to go through, once mono-tagged forms cause a decrease in the number of possible combinatories, reducing also its time of search

ExtracTerm: learning rules

Linguistic learning rules are build from the analysis of the lexical context of the form to be disambiguated. Each time ExtracTerm identifies a grammatical class to be disambiguated, it searches left and right of the form and applies the rules, opting for the tag suitable to the context in question.

Rules may thus be established positively. When condition A and B are true, situation C is obtained: [+A] + [+B] = [C]. In given contexts it is more productive to establish rules negatively. This way, when conditions A and B are true, condition C is not obtained: [+A] + [+B] = [-C]:

Rule

[Art:def:f:s] [Pron:pess:f:s] [Pron:dem:f:s] [Prep1:a] + [N:f:s] → [Art:def:f:s] + [N:f:s]

Each time form *a* is tagged with [Art:def:f:s] [Pron:pess:f:s] [Pron:de,:f:s] [Prep1:a] and the grammatical class to its *right* is followed by tag [N:*], then ExtracTerm will chose tag [Art:def:f:s], in detriment of the others.

Example: a [Art:def:f:s] imagem [N:f:s]

With the results obtained from the application of these five learning rules, that allow the disambiguation of form *a* in five different contexts, we believe to be able to test matrix N* + Prep1:a + N*

Then, the order in which the rules are the following: a) Tagging the *corpus*: it corresponds to the attribution of all meta-linguistic tags a given form may assume, independently of the context in which it occurs; b) Applying disambiguation rules: ExtracTerm applies the disambiguation rules, whose aim is that of abolishing multi-tags, in order to proceed into the applying of the following rules; c) Applying learning rules: applying these rules consists of identifying the pre-defined structures that multilexemic terminological units may assume; d) Extraction of multilexemic terminological units

ExtracTerm: an extractor

Based on the tagged texts, ExtracTerm applies the matrixes contained on a dictionary of typologies. To exemplify, we observed the behaviour of the base forms tagged with [N] and [Sigla].

ExtractTerm carries out the consecutive extraction of each form contained in the corpus, attached by the meta-linguistic tag [N] or [Sigla]. These two tags are the starting-points for the program to begin the process underlying extraction.

This being so, the extractor begins its course, stopping at the first [N] or [Sigla] it finds. It checks if the tag sequence to the right of [N] or [Sigla] corresponds to the first of the 666 rules considered in the dictionary.

If the search is successful, ExtracTerm will extract the sequence, followed by the matrix corresponding to its structure, to an HTML file. If it does not find a corresponding structure, it will go straight into the following rule, until it looks at all the matrixes. This operation is repeated, sequentially, for all [N] and [Sigla] tags, from strong punctuation to strong punctuation.

Based on the tagged *corpus*, ExtracTerm will carry out the following extraction:

1. Teledetecção [N:f:s] em [Prep1:em] áreas [N:f:pl] {N:f:s + Prep1:em + N:f:pl}
2. áreas [N:f:pl] periurbanas [Adj:f:pl] {N:f:pl + Adj:f:pl}
3. combinação [N:f:s] de [Prep1:de] índices [N:m:pl] {N:f:s + Prep1:de + N:m:pl}
4. combinação [N:f:s] de [Prep1:de] índices [N:m:pl] temáticos [Adj:m:pl] {N:f:s + Prep1:de + N:m:pl + Adj:m:pl}
5. índices [N:m:pl] temáticos [Adj:m:pl] {N:m:pl + Adj:m:pl}
6. mudança [N:f:s] de [Prep1:de] uso [N:m:s] [V:cj] {N:f:s + Prep1:de + N:m:s}
7. uso [N:m:s] [V:cj] do [Prep2:do:m:s] solo [N:m:s] {N:m:s + Prep2:do:m:s + N:m:s}
8. solo [N:m:s] com [Prep1:com] recurso [N:m:s] {N:m:s + Prep1:com + N:m:s}
9. imagens [N:f:pl] digitais [Adj:2gen:pl] {N:f:pl + Adj:2gen:pl}
10. imagens [N:f:pl] digitais [Adj:2gen:pl] SPOT [Sigla] {N:f:pl + Adj:2gen:pl + Sigla}
11. imagens [N:f:pl] digitais [Adj:2gen:pl] SPOT [Sigla] HRV [Sigla] {N:f:pl + Adj:2gen:pl + Sigla + Sigla}
12. SPOT [Sigla] HRV [Sigla] {Sigla + Sigla}

The obtained results allow us to go on with its analyses, which aim is to explain the dynamic functioning of ExtracTerm.

ExtracTerm has compared the corpus to the dictionary of typologies. Each time he finds to the right of [N] a tag corresponding to sequences recorded on the dictionary of typologies, he effectuates an extraction. That was the case with *teledetecção* [N], with *combinação* [N], but not with the noun *recurso* [N], because the sequence N + Prep2:às has not been foreseen inside the typology. For the remaining [N], ExtracTerm has extracted all sequences foreseen on the dictionary of typologies.

Let us have a look at sequences (1), (2) and (3), corresponding to utterance *combinação de índices temáticos*:

In order to extract the maximum sequence *combinação de indices temáticos*, we have applied a complex rule {N:f:s + Prep1:de + N:m:pl + Adj:m:pl}, that results from two elemental rules {N:f:s + prep1:de + N:m:pl} and {N:m:pl + Adj:m:pl}.

In such cases, we have noticed that rule {N:f:s + Prep1:de + N:m:pl + Adj:m:pl} is dispensable, once the phraseology is recovered with first and second rules. The outcome of this option is the following:

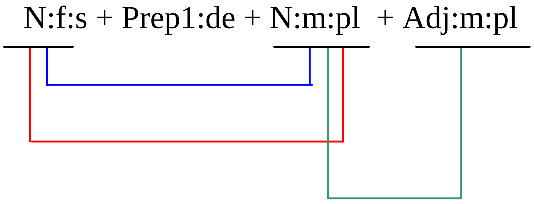
- combinação de índices {N:f:s + Prep1:de + N:m:p}
- combinação de índices temáticos {N:f:s + Prep1:de + N:m:pl + Adj:m:pl}
- índices temáticos {N:m:pl + Adj:m:pl}

Recognition of the phraseology results from the fact that the last [N] of the first rule coincides with the first [N] of the second rule, obtaining the sequence *combinação de índices índices temático*, in which one of the indexes [N] is eliminated at the time of the interpretation of the context involving [N], that we shall designate as lexical context.

Nevertheless, we prefer to keep the three rules, regardless of the clear redundancy in rule {N:f:s + Prep1:de + N:m:pl + Adj:m:pl} and even though this rules correspond not to a multilexemic terminological unit, but to a phraseology. In our understanding, it is preferable that the outcome of the extraction contains more noise and less silence.

Whereas the first extraction, *combinação de indices* {N:f:s + Prep1:de + N:m:p}, must be ignored, because it corresponds to a 'terminological free sequence', applying the second rule has lead to the extraction of a phraseology consisting of the lexeme *combinação*, generic term of scientific orientation, and the multlexemic terminological units *índices temáticos*. We consider, therefore, the phraseology a minimum context of the multlexemic terminological units, that is, a lexical context.

The components of the phraseology may be represented thus:



From the three following sequences

(4) mudança de uso {N:f:s + Prep1:de + N:m:s}
(5) uso do solo {N:m:s + Prep2:do:m:s + N:m:s}
(6) solo com recurso {N:m:s + Prep1:com + N:m:s}

Sequences (4) and (5) are excluded *per se*, in as much as we have classified them as 'terminological free sequences'.

However, the overlapping of the matirx applied in (5) with the matrix applied in (6) allows us to arrive to an interpretation close to the previous one: *mudança de uso de solo* is phraseology from with the term *uso do solo* can be extracted:

N:f:s + Prep1:de + N:m:s + Prep2:do + N:m:s ~~+ Prep1:com + N:m:s~~

In this case, the first rule withholds *mudança de uso* and the second rule withholds the sequence *uso do solo*. It is the combination of both rules that allows the isolation of the terminological unit, the [N] link between the two rules being the term *uso*: N:f:s + Prep1:de + N:m:s + prep2:do + N:m:s.

Sequences *mudança de uso* and *combinação de indices* are 'terminological free sequences', because the sequence of its lexemes corresponds to a pertinent cut in the text, there being, however, no lexical combinations to render them solidary. *Mudança de uso* and *combinação de indices* both have frequency 1 and occur in contexts where their specifiers, placed to their right, are omitted, because they have been largely referenced to previously.

The same does not happen to *solo com recurso*. In this case, structure N:m:s + Prep1:com + N:m:s produces a misleading split of two cohesive syntactic sequences: terminological unit *uso do solo* and prepositional locutions *com recurso a*.

Such has taken place because ExtracTerm must stop at *solo* [N] and realise that the expansion to the right corresponds to a pre-registered matrix. The mistake is due to two factors mainly.

The first is related to the fact of us having included the preposition *com* on the list of grammatical morphemes pertinent for the constituent of multilexemic terminological units. From observing the results, we have noticed that this preposition is not a constituent preposition, but a preposition whose function is that of associating two terminological units, as we can prove with the following examples:

[área$_{N}$ cartografada $_{Adj}$]$_{N}$ com [lissage$_{N}$] $_{N}$
[ensaios$_{N}$] $_{N}$ com [filmes$_{N}$ infravermelhos$_{Adj}$] $_{N}$
[píxeis$_{N}$] $_{N}$ com [valores$_{N}$ radiométricos$_{Adj}$] $_{N}$
[observação$_{N}$ da$_{Prep}$ terra$_{N}$] $_{N}$ com [imagens$_{N}$ numéricas$_{Adj}$ multiespectrais$_{Adj}$]$_{N}$

Preposition "*com*" has the function of associating on the syntagmatic level two terms both endowed with semantic autonomy.

The second factor concerns the non-inclusion of the locution *com recurso a* on the dictionary of locution, what has lead EtiqueLex into tagging *com* [Prep1:com] *recurso* [N:mf:s] *a* [Art:f:s] [Pron:pess:3p:f:s] [Pron:dem:3p:f:s] [Prep1:a], ExtractTerm acting therefore on the prepositional locution as if it was a free sequence.

As a result from such an analysis, we have added the locution *com recurso* a to the dictionary of locution and have deleted matrixes holding the tag [Prep1:com], because the noise provoked by such matrix is higher than the silence provoked by its absence.

The last four sequences are examples of an extraction well succeeded:

(7) imagens digitais {N:f:pl + Adj:2gen:pl}
(8)imagens digitais SPOT {N:f:pl + Adj:2gen:pl + Sigla}
(9) imagens digitais SPOT HRV {N:f:pl + Adj:2gen:pl + Sigla + Sigla}
(10) SPOT HRV {Sigla + Sigla}

In this case, the four examples correspond to four multilexemic terminological units, once they correspond to four distinct denominations**,** and the expansions to the right of (8) and (9) come to specify the type of *imagens digitais*, ExtracTerm having, on these cases, functioned with its maximum potentiality:

[[[imagens $_{N}$ digitais $_{Adj}$] $_{N}$ [[SPOT $_{Sigla}$] $_{N}$ HRV $_{Sigla}$] $_{N}$]$_{N}$

If we understand that *imagens digitais* plays the role of a noun, we can, by extension, consider that *[SPOT]* and *[HRV]* play the role of 'epithet acronyms' at the same level of epithet nouns.

For the specialist, the adjective digital may be omitted, without establishing conceptual ambiguity. In fact, all satellite images are, nowadays, digital (means of registering the data), and the term *imagens SPOT* may occur, the acronym *[SPOT]*, name of the French satellite, qualifying the noun [imagens digitais]*N* or the noun [imagens]*N*.

On the other hand, acronym *[HRV*], meaning *Haute Résolution dans le Visible*, is an image acquisition system inherent to *satélite SPOT*, what allows us to infer that the acronym *[HRV]* plays the role of a qualifying adjective, and must therefore be sub-categorized by a noun, and in truth that is the case with acronym [*SPOT*].

Sequence *SPOT HRV* is a terminological unit and may consequently be classified as a noun, what allows us to give it the function of epithet noun, at the same level of the acronym *[HRV]*, once it unequivocally qualifies the term [imagens N digitais Adj]*N*.

The correct extraction of such terms leads us into concluding that the rules have good applicability under given conditions, what compels us into improving them.

Conclusion

With ExtacTerm we have extracted collocations, multlexemic terminological units and phraseologies.

However, we have also extracted less interesting sequences, such as current 'terminological free sequences' and current multilexemic units, as well as sequences resulting from badly carried out cuts.

The ExtracTerm ia a dynamic tools which permit to alter- modifying and/or adding - morphosyntactical conceptual and semantic descriptions. These tools also permit to increase opportunely, the terminogénica structures, and the recognition rules, as well as the learning rules when the systematic observation of special purposes language *corpora* is being done.

Bibliography

- Bourigault, Didier (2001), "Introduction", *Recent Advances in Computational Terminology*, Ed. by Didier Bourigault, Christian Jacquemin, Marie-Claude l'Homme, Amsterdam / Philadelphia, John Benjamins.
- Costa, Rute (2001), *Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas*, Dissertação de Doutoramento, Universidade Nova de Lisboa.
- Costa, Rute, Silva, Raquel (2004), "The verb in terminological collocation. Contribution to the Development of a Morphological Analyser – MorphoComp", *Proceedings of LREC 2004*, Lisbon.
- Heid. U. (2001), Collocations in Sublanguage Texts : Extraction from Corpora, In Handbook of Terminology, Volume 2, compiled by Sue Allen Wright, Gerhard Budin, Amsterdam / Philadelphia, John Benjamins.

Database for Evaluating Extracted Terms and Tool for Visualizing the Terms

Masaki Murata*, **Qing Ma**†, **Tamotsu Shirado***, **Hitoshi Isahara***

*National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{murata,shirado,isahara}@nict.go.jp

†Ryukoku University
Otsu, Shiga, 520-2194, Japan
qma@math.ryukoku.ac.jp

Abstract

We constructed a database that can be used to evaluate term extraction. It can be used to calculate precision and recall rates because it is very exhaustive. We did experiments on term extraction and compared various kinds of methods. We also applied automatic application tools to extract terms that displayed the results in the form of a two-dimensional figure in about 20 seconds.

1. Introduction

Term-extraction techniques are used in various domains, and various kinds of terms are extracted. Recently, for example, there have been studies on open-domain question-answering systems where many kinds of terms (e.g. names of people, countries, and mountains.), used as knowledge to answer questions related to such terms, have been extracted (Murata et al., 2002). There have also been studies on bio-informatics where protein names used as terms (Mitsumori et al., 2004) have been extracted. We therefore created a database that could be used to evaluate the results of automatic term extraction in different domains. The database contains about 60 categories (e.g., names of countries, cities, and planets.). We also constructed a term-pair database with about 60 categories. It contains pairs of terms such as the name of a country and its capital city. These databases include almost all the terms in each category. Thus, they can be used to calculate the precision and recall rates of term extraction. They contain a total of about 30,000 terms and 100,000 data items. They also contain lists of synonyms for each term. In this paper, we explain the database and experiments we did using the database. Moreover, we describe automatic application tools for extracting terms and displaying the results. The tools are very useful and can output a figure that contains numerous terms in about 20 seconds

2. Construction of database to evaluate automatic term extraction

We constructed a database for this study to evaluate automatic term extraction in the Japanese language. There are examples of entries in the database in Tables 1 and 2. Table 1 is a term list for the names of countries. Each line has the name of a different country. The first item is a basic form for the entry in the line and its different forms are stored at the latter part of the same line. Table 2 is a term pair list for pairs of names for countries and their capitals. Each line has the name of a country and its capital, the same as in Table 1. In this paper, a list such as that in Table 1 is called a *one-term list* and one such as that in Table 2 is called a *two-term list*. We prepared 58 lists each for the one-term and two-terms. The amount of data for the one-term and two-terms list database is listed in Table 3. The constructed lists include “planets in the solar system”, “satellites in the solar system”, “Japanese national holidays”, “space shuttles”, “names of flowers”, “positions in a soccer game”, “names of Japanese professional baseball players” and “world heritage sites”. They also include “pairs of names of villages and their prefectures”, “pairs of Japanese national holidays and their dates”, “pairs of planets and their satellites in the solar system”, and “pairs of composers and their music”.

Table 1: Example of one-term list (names of countries)

Iceland, the Republic of Iceland, ISL
Ireland, the Republic of Ireland, IRL
Azerbaidzhan, the Republic of Azerbaidzhan, AZE
Azores
Adygeiskaya, the Republic of Adygeiskaya
Afghanistan, the Republic of Afghanistan
America, the United States of America, USA
...

How we constructed the database is described below:

- A term list was constructed by consulting a dictionary or a guide book (e.g., for planets in the solar system, satellites in the solar system, Japanese national holidays, or space shuttles).
- A term list was constructed by consulting a web site (e.g., for names of dramas on Japanese TV).
- A term list was constructed by consulting and combining a dictionary, a guide book, and a web site (e.g., for names of mountains in the world and the names of flowers).
- A term list was constructed by an expert with a rich knowledge of a related domain (e.g., for positions in soccer).

How we constructed the different forms is explained below:

Table 2: Examples of two-term list (pairs of names of countries and their capitals)

Iceland, the Republic of Iceland, ISL	Reykjavik
Ireland, the Republic of Ireland, IRL	Dublin
Azerbaidzhan, the Republic of Azerbaidzhan, AZE	Baku
Azores	Angra
Adygeiskaya, the Republic of Adygeiskaya	Maikop
Afghanistan, the Republic of Afghanistan	Kabul
America, the United States of America, USA	Washington, Washington D.C., Washington DC
...	

Table 3: Size of databases

	one-term list database	two-term list database
Number of lists	58	58
Number of basic forms	17696	19387
Number of basic forms and different forms	26728	106850

- Different forms with regularities were generated through regularities automatically (e.g., a surname was extracted from a person's name).

- Different forms for each term were extracted from a web site (e.g., names of mountains in the world and names of flowers).

- Different forms for each term were generated by a person to the best of their ability (e.g., names of constellations).

- Different forms for each term were constructed by an expert with a rich knowledge of a related domain (e.g., positions in soccer).

How each term list and different forms in each term list were constructed, the definitions of a basic form and different forms in each term list, comments from the person constructing each term list, and information on the degree of completeness of each term list are described in the database. For example, the definition of a basic form in the term list for "countries' names" is "the basic form is not a formal form but the form that is used most frequently (e.g., the basic form for "the French Republic" is "France").

For information on the degree of completeness of each term list, "completed almost 100%", "not completed", and others are described. The "completed almost 100%" means a term list with almost all its members stored in it. The "not completed" means a term list with some of its members not stored in it. We used entries whose members were limited such as countries' and capitals' names as term lists, so we could construct term lists with almost all members stored in them. Therefore, the database we constructed for this study could be used to calculate the precision and recall rates for experiments on automatic term extraction. There were 58 categories in the term lists. This is fewer than in term lists throughout the world. By specifying a specific domain for terms and decreasing the kinds of term lists, we could construct term lists that were almost complete and that could be used to calculate precision and recall rates. The database constructed for this study can be used to evaluate any kind of data that is used as the target for extraction. For example, we can evaluate results extracted from newspaper articles and we can evaluate results extracted from the Web with our database.

A two-term list consisting of pairs of terms could be considered as knowledge. Recently, studies have been done that have considered question-answering to be important. If a question-answering system has a two-term-list database to answer questions, it performs better (Fleischman et al., 2003). Our two-term-list database can be applied to such question-answering systems and can also be used to evaluate data in systems extracting two-term lists (Brin, 1998).

3. Experiments on automatic term extraction from newspaper articles and their evaluation based on our database

We conducted brief experiments on automatic term extraction and evaluated the results with the database we constructed. We used "completed almost 100%" term lists with 40 categories for the one-term and 44 categories for the two-term list database. We used a method where the system learns very few positive examples and then extracts more positive examples based on the learning results. A different form for each term was not considered to be positive and only a basic form was considered to be positive in the experiments. We used five basic forms that frequently occurred in the Japanese Mainichi newspaper as the input for the few positive examples. We used the Japanese Mainichi newspaper (1991-2000) as the target data D for extracting terms. The algorithm for extracting terms is as follows:

1. The system searches a positive input example in the D. It extracts a pattern occurring near plural positive input examples as c_i (near pattern will be defined later.)

2. It then searches c_i in D. It extracts expressions as exp using c_i.

Table 4: Experimental results on one-term-list database

	No use of character kinds and KR			Use of character kinds			Use of KR			Use of character kinds and KR		
	AP	RP	TP	AP	RP	TP	AP	RP	TP	AP	RP	TP
Method 1	0.102	0.170	0.305	0.142	0.220	0.365	0.096	0.161	0.255	0.154	0.231	0.360
Method 2	0.174	0.235	0.445	0.182	0.244	0.475	0.177	0.235	0.465	0.185	0.247	0.490
Method 3	0.171	0.234	0.435	0.178	0.242	0.460	0.182	0.239	0.475	0.189	0.251	0.490
Method 4	0.172	0.234	0.435	0.179	0.244	0.465	0.172	0.235	0.435	0.179	0.245	0.465
Method 5	0.174	0.236	0.410	0.192	0.264	0.450	0.190	0.246	0.460	0.206	0.272	0.490

3. Finally, it sorts the extracted exps in descending order of $Score$ value and outputs them as terms.

We used the following equations as $Score$.

- Method 1 (decision-list method)

$$Score = max_i p_i \tag{1}$$

- Method 2 (Simple-Bayes method)

$$Score = \prod_i p_i \tag{2}$$

- Method 3 (method based on similarities)

$$Score = \sum_i 1 \tag{3}$$

- Method 4 (a previous study (Riloff and Jones, 1999))

$$Score = \sum_i (1 + 0.01 p_i log(f_i)) \tag{4}$$

- Method 5 (a previous study (Murata and Isahara, 2002))

$$Score = 1 - \prod_i (1 - p_i), \tag{5}$$

where f_i is the number of positive inputs when c_i occurs nearby and p_i is the ratio of positive inputs in the expressions extracted by c_i. In Methods 1, 2, 4, and 5, when $Score$ has the same value, its value in Method 3 is used to sort exps. In Method 3, when $Score$ has the same value, its value in Method 5 is used to sort exps.

We conducted experiments with our one-term-list database. We used 1 to 3 gram characters with the character just at the left or at the beginning of a positive input and 1 to 3 grams characters with the character at the end or at the right of a positive input as the pattern c_i. The experimental results are listed in Table 4. AP is the average precision and RP is the r-precision used in information retrieval (Murata et al., 2000). TP is the average precision in the top-five output terms. We used two methods, using character kinds and KR. In the first method, we did not extract expressions, including character kinds, that were not included as positive input.[1] The second KR method uses the ratio of positive input where c_i occurs nearby. Here, $p_i * f_i / n_i$ is used instead of $p_i$to calculate the value of $Score$. In Method 3, f_i is used instead of 1. n_i is the number of the positive input examples. We eliminated different forms of positive examples when we evaluated the extracted results.

There have been some previous studies (Method 4) (Riloff, 1996; Riloff and Jones, 1999).[2] We compared these various methods or equations from the previous studies using our evaluation database. We found Method 5, using character kinds and KR, was best. We next conducted experiments using the two-term-list database. We also used five-term pairs as positive input. We used combinations of the order of the two terms of a positive input, i.e., 1 to 3 gram characters with the character just at the left or beginning of the first term of a positive input and an expression between the two terms of a positive input and combinations of the order of the two terms of a positive input as the pattern c_i. We also used 1 to 3 grams characters with the character just at the right or end of the second term of an input positive and an expression between the two terms of a positive input as the pattern c_i. When the left or right side of an extracted exp was not restricted by c_i, it was divided and extracted using a Japanese hiragana character (a functional word) to define its boundary. In the experiments using the two-term list, Method 3 with character kinds and KR was best. AP was 0.026, RP was 0.040, and TP was 0.141.

4. Automatic application tools for extracting terms and displaying results

Because the method in the previous section uses a high-speed string search algorithm such as a suffix array, it has the advantage of extracting terms very quickly. Consequently, we developed a simple application tool using this. First, the user inputs a few words. The tool uses them as positive input and gathers terms that are in the same domain based on Method 5 using character kinds and KR. We simplified c_i and reduced the amount of data to speed up the process. As the tool visually displays extracted terms in the same category, they can be seen easily. We used SOM (Ma et al., 2002) for the visualization and reduced the number of

[1] The Japanese language has several kinds of characters such as Chinese, Hiragana, and Katakana.

[2] The boot-strapping algorithm was used in the previous studies, in addition, and they obtained more precise terms by using more positive examples. Our precisions could have been improved had we used this.

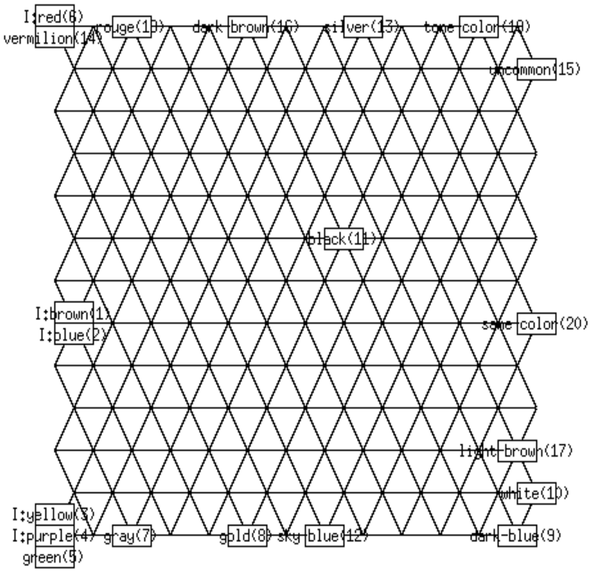


Figure 1: Term extraction for “color” input by user

learning iterations to speed it up. We used the same method as (Ma et al., 2002) and used c_is except for those using inner expressions of positive input as the context for SOM. The tool could output a figure containing numerous terms in about 20 seconds. (It was partly developed with Perl and speed can be improved by using C instead of Perl or modifying the algorithms.) The tool obtained 0.111, 0.164, and 0.310 for AP, RP, and TP. The precision was decreased because some information had to be eliminated due to speeding it up.

Figure 1 shows the output results when a user inputs “red”, “blue”, “yellow”, “purple”, or “brown” to the tool. Positive input terms are labeled “I:”. The number attached to each term indicates in what order each item was extracted. In general, using SOM, as expressions with a similar meaning are located near to one another, the resulting figure can be seen more easily. In this example, “vermilion” and “rouge” are located near “red” at the left and upper corner. We found that similar terms were located near each other. The “tone-color”, “uncommon”, and “same-color” were located near one another at the right and upper corner. These terms were not colors. Visualization has a function that gathers such unnecessary terms in the one place.

5. Conclusion

We constructed a database that could be used to evaluate term extraction. It contained a one-term-list database and a two-term list database with “countries’ names” and “pairs made up of their names and capitals” for the evaluation. As it is very exhaustive, it can be used to calculate precision and recall rates. A two-term list consisting of pairs of terms is considered as knowledge and it can be used for studies on knowledge processing such as in question-answering systems. We did experiments on term extraction for this study and compared various kinds of methods. Method 5 using character kinds and KR was best in the one-term-list database, and Method 3 using character kinds and KR was best in the two-term-list database. We discussed automatic application tools for extracting terms that displayed the results in the form of a two-dimensional figure in about 20 seconds. These demonstrated two advantages in that similar terms were located near each other and unnecessary terms were collected in other locations.

6. References

Brin, Sergey, 1998. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT’98*.

Fleischman, Michael, Eduard Hovy, and Abdessamad Echihabi, 2003. Offline strategies for online question answering: Answering questions before they are asked. In Erhard Hinrichs and Dan Roth (eds.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Ma, Qing, Min Zhang, Masaki Murata, Ming Zhou, and Hitoshi Isahara, 2002. Self-organizing Chinese and Japanese semantic maps. *International Conference on Computational Linguistics (COLING’2002)*:605–611.

Mitsumori, Tomohiro, Sevrani Fation, Masaki Murata, Kouichi Doi, and Hirohumi Doi, 2004. Boundary correction of protein names adapting heuristic rules. In *Fifth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2004)*.

Murata, Masaki and Hitoshi Isahara, 2002. Automatic extraction of differences between spoken and written languages, and automatic translation from the written to the spoken language. In *LERC 2002*.

Murata, Masaki, Kiyotaka Uchimoto, Hiromi Ozaku, Qing Ma, Masao Utiyama, and Hitoshi Isahara, 2000. Japanese probabilistic information retrieval using location and category information. *The Fifth International Workshop on Information Retrieval with Asian Languages*:81–88.

Murata, Masaki, Masao Utiyama, and Hitoshi Isahara, 2002. A question-answering system using unit estimation and probabilistic near-terms IR. *Proceedings of the Third NTCIR Workshop (QAC)*.

Riloff, Ellen, 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of AAAI-96*.

Riloff, Ellen and Rosie Jones, 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI-99*.

English Terminology in CLAT

Michael Carl[1]**, Maryline Hernandez**[1]**, Susanne Preuß**[1] **and Chantal Enguehard**[2]

[1]Institut für Angewandte Informationsforschung
Martin-Luther-Str. 14, Saarbrücken, Germany
{carl,maryline,susannep}@iai.uni-sb.de

[2]Institut de Recherche en Informatique de Nantes
2, rue de la Houssinière, 44322 Nantes, Cedex 3 France
Chantal Enguehard@irin.univ-nantes.fr

Abstract

CLAT (Controlled Language Authoring Technology) is a tool that supports authors in the production of technical documents. Much work is currently invested to enhance the linguistic components for the English version of CLAT. In this paper we give a short overview of CLAT. We introduce the terminology tool, its underlying technology and describe the graphical interface in CLAT. The paper summarizes an experiment showing figures of precision and recall and discusses future developments.

1. Introduction

CLAT (Controlled Language Authoring Technology) is a tool designed to support technical authors in producing and revising technical documents in various domains. The German version of CLAT (i.e. MULTILINT) is most advanced and already in use in a number of companies (Haller et al., 2002)[1]. Much work is currently invested in its English version to enhance the checking possibilities for grammar and terminology.

In this paper we report on ongoing work to extend the terminology component of CLAT for English. In order to effectively check compliance to corporate language requirements, the CLAT term base encodes information concerning the status of the terms: preferred form, admitted form or deprecated form. Moreover, on the basis of morpho-syntactic analysis and lemmatisation, the tool is able to detect typographical and derivational variants in texts.

Using an abductive mechanism (Carl et al., 2002; Carl et al., 2004) the existing terminology is extended with term variation templates. These templates are produced by combining general knowledge of variation patterns with the terminological entries. Term templates can be further enriched with synonymy relations. When checking terminological consistency, CLAT matches a document against the extended terminology and marks the detected terms or term variants. In case the status of the matched entry is other than preferred or admitted, the term variant or the deprecated term is highlighted and the author is provided with a message and the preferred base form.

While this technique is described in detail in (Carl et al., 2002; Carl et al., 2004), this paper presents the graphical interface of CLAT. In section 2. we give an overview over CLAT and its GUI. Section 3. discusses the terminology component in more detail. Section 4. outlines an experiment and section 5. discusses the results.

2. Controlled-Language Authoring Technology

Controlled-Language Authoring Technology (CLAT) has been developed to suit the need of some companies to automatically check their technical texts for general language and corporate language conventions. Within CLAT, texts are checked with respect to:

- orthographic correctness
- company specific terminology and abbreviations
- general and company specific grammatical correctness
- stylistic correctness according to general and company specific requirements

The orthographic control examines texts for spelling mistakes and proposes alternative writings. The terminology component matches the text against a terminology and abbreviation database where also term variants are detected. The grammar control

[1]see also `http://iai.uni-sb.de/iaide/de/clat.htm`

checks the text for grammatical correctness and disambiguates multiple readings while stylistic control detects stylistic weaknesses.

The components build up on each other's output. Their modularity is suited to adapt to different texts and requirements. Besides the described control mechanisms, CLAT also has a graphical frontend as shown in figure 1. The lower part in figure 1 plots an input paragraph while the upper part shows the automatically annotated paragraph with mistakes highlighted. The document structure is shown in the left part of the window. A user can revise a document by clicking through the paragraph symbols. The linguistic engine works in the background performing morphological analysis, lemmatisation, terminology checking, shallow parsing and style control while the GUI marks erroneous segments in the texts with different colors.

The user can click on one of the automatically annotated errors in the upper part of the window to display an explanatory message in the middle part of the screen. A separation is made between mistakes on the word level, on a grammatical level and on a stylistic level. The user can switch between the different levels of analysis by clicking on the appropriate buttons. The text in the lower part can also be edited, re-checked and stored.

3. The Terminology Tool in CLAT

As a general rule and to enhance readability, terms in technical documents should be used consistently and in accordance with an authorized terminology. However, people often use different linguistic forms to name the same thing. To cope with this problem, CLAT tries to anticipate and detect variants of terms. In section 3.1. we give a background of the implementation while section 3.2. shows a segment of the graphical interface.

3.1. Architecture

There are basically two ways of matching a candidate variant in a document (or in a list of terms) onto a list of authorized terms: a database approach and a run-time approach:

1. in the run-time approach, a candidate sequence of words in the document undergoes a number of transformations which map it onto an authorized term. The original sequence in the document is then marked as a variant of the authorized retrieved term.

2. in the database approach a limited number of possible variants are generated for each authorized term. The variants are stored in a database with a link to their authorized terms. A matched sequence of words in the document is marked as a variant of the term from which the database entry was generated.

CLAT's terminology tool implements a database approach. The database approach has the drawback that all possible variants which the tool is supposed to recognize are need be generated and stored in a database. However, since CLAT can store underspecified variants the size of the base only marginally increases compared to the gain of coverage. The outstanding advantage of the database approach is, however, *log*-time retrieval.

The run-time solution transforms and maps a candidate sequence of words in a document onto its authorized forms in the database. Time required for this mapping increases linearly in time with every possible transformation that the sequence in the document undergoes. Jacquemin's *FASTR* (Jacquemin, 2001) implements the run-time approach. Using a set of metarules, Jacquemin remains below this linear limit.

CLAT's terminology tool integrates a rule-based approach and an example-based approach. Rules are used to generate variation templates from authorized terms which are stored in a database. Rules are also used to consolidate the findings of the matching process. The technique underlying this process is described in-depth in (Carl et al., 2004; Carl et al., 2002).

3.2. Graphical Interface

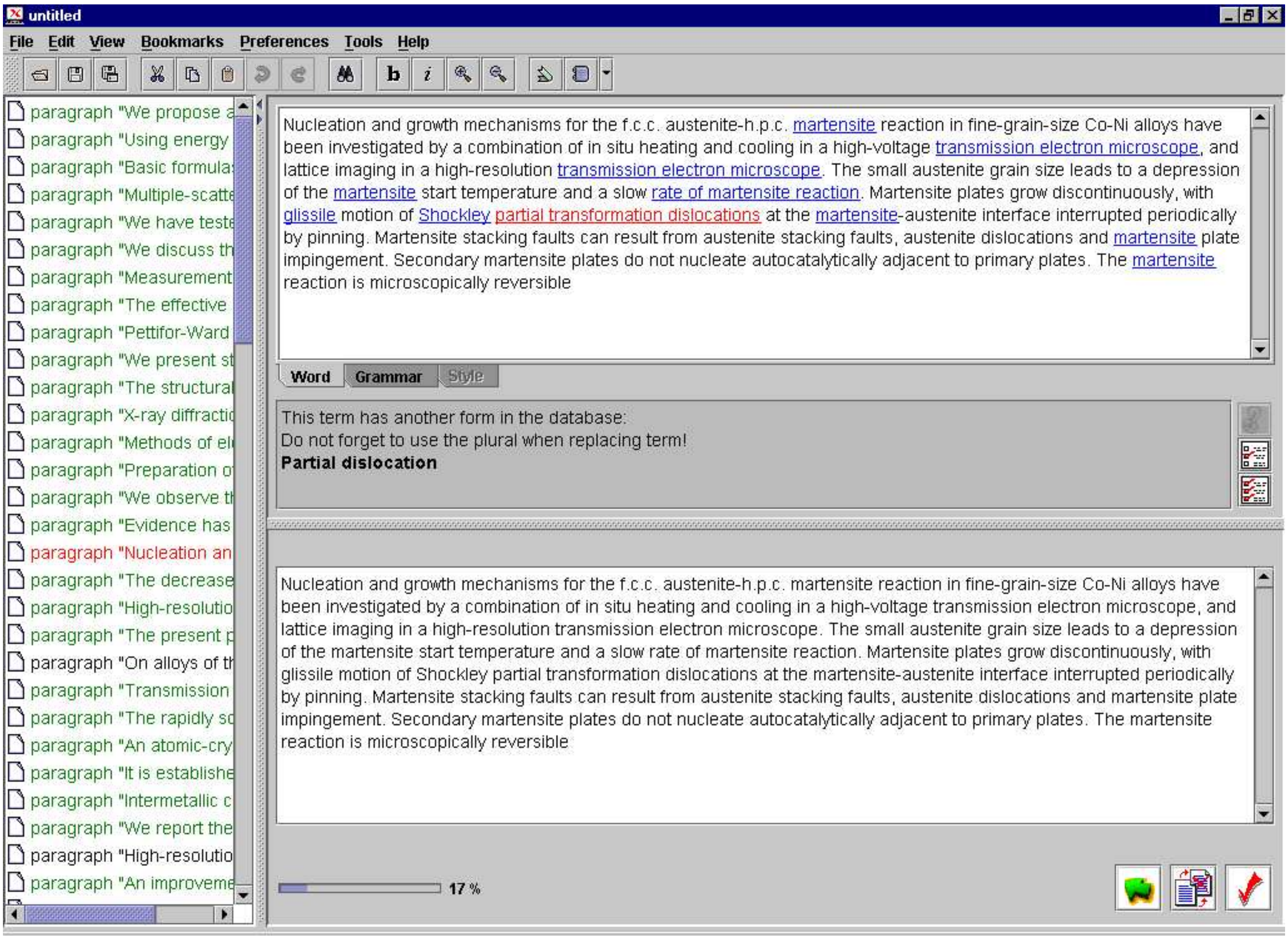
Figure 1 highlights mistakes on the word level. Two types of mistakes are differentiated on the word level: either the word (or word form) cannot be analyzed or is unknown to the system, or it is detected a deprecated form or a term variant.

The word "martensite" is unknown to the system. It is not in the system's dictionary nor in the terminology database. By clicking on one of the instances, the user is prompted the following message in the middle window:

> This word is unknown. Does it contain a spelling mistake?

The user can click on an ignore button for the word to be admitted in the paragraph or for the entire document.

There is a further occurrence of "martensite" within the compound noun "rate of martensite reaction". This compound is detected a permutation variant of the term "Reaction rate". The user is prompted the message:



Nucleation and growth mechanisms for the f.c.c. austenite-h.p.c. martensite reaction in fine-grain-size Co-Ni alloys have been investigated by a combination of in situ heating and cooling in a high-voltage transmission electron microscope, and lattice imaging in a high-resolution transmission electron microscope. The small austenite grain size leads to a depression of the martensite start temperature and a slow rate of martensite reaction. Martensite plates grow discontinuously, with glissile motion of Shockley partial transformation dislocations at the martensite-austenite interface interrupted periodically by pinning. Martensite stacking faults can result from austenite stacking faults, austenite dislocations and martensite plate impingement. Secondary martensite plates do not nucleate autocatalytically adjacent to primary plates. The martensite reaction is microscopically reversible

This term has another form in the database:
Do not forget to use the plural when replacing term!
Partial dislocation

Nucleation and growth mechanisms for the f.c.c. austenite-h.p.c. martensite reaction in fine-grain-size Co-Ni alloys have been investigated by a combination of in situ heating and cooling in a high-voltage transmission electron microscope, and lattice imaging in a high-resolution transmission electron microscope. The small austenite grain size leads to a depression of the martensite start temperature and a slow rate of martensite reaction. Martensite plates grow discontinuously, with glissile motion of Shockley partial transformation dislocations at the martensite-austenite interface interrupted periodically by pinning. Martensite stacking faults can result from austenite stacking faults, austenite dislocations and martensite plate impingement. Secondary martensite plates do not nucleate autocatalytically adjacent to primary plates. The martensite reaction is microscopically reversible

Figure 1: Terminology Checking in CLAT

> This term has another form in the database:
> **Reaction rate**

Two occurrences of “transmission electron microscope” are detected as derivation variants of “Transmission electron microscopy”. By clicking on one of the instances, the following message appears in the middle window:

> This term has another form in the database:
> **Transmission electron microscopy**

Due to linguistic analysis (Maas, 1996; Maas, 1995), CLAT also detects different inflections in the authorized term and the variant occurrences in the text. For instance “partial transformation dislocations” is detected an insertion variant of the authorized term “Partial dislocation”. By clicking on the marked variant in the upper window, the following message appears in the middle of the screen:

> This term has another form in the database:
> Do not forget to use plural when replacing term
> **Partial dislocation**

Since the variant in the text shows a different agreement in number than the authorized form in the database, the author is hinted to adjust to plural when replacing the term.

Some variants can be traced back to a number of authorized terms. The term “state structure”, for instance, is the head of two authorized term, “Liquid state structure” and “Amorphous state structure”. The author is prompted the following message:

> This term has another form in the database:
> **Liquid state structure**
> **Amorphous state structure**

4. Experiment

In an experiment we evaluated and compared the terminology tool based on a collection of 1280 abstracts on the chemistry of metals. Terms and variants of terms in the abstracts were manually annotated in a previous project (Enguehard, 2003). A terminology database was provided containing 6602 authorized base terms.

	Fastr	Syrete	*T*	*TV*
Recall	64.63	70.44	66.43	68.42
Precision	89.05	98.81	96.70	94.33

Table 1: Recall and Precision for Syrete, FASTR and CLAT

The abstracts were automatically annotated using three different term recognition systems: Syrete[2], *FASTR* (Jacquemin, 2001) and CLAT's terminology tool. Two versions of the CLAT's terminology tool were used. In the version *T* only the lexemes of the 6602 authorized base terms were indexed. The version *TV* contained the 6602 terms plus 22367 variation templates. These variation templates were generated from 12 variation patterns to detect reduction, insertion and permutation variants (see (Carl et al., 2004) for a similar experiment) so that for each base term on average four variation templates were indexed. Results of the four experiment in terms of precision and recall are shown in table 1.

5. Discussion

Despite the excellent results in table 1 for all four settings, a number of open questions remains.

For example, the compound noun "rate of martensite reaction" was annotated a variant of "Reaction rate" in the test text by a specialist in the domain (Enguehard, 2003). However, variants built by several variation mechanisms such as permutation and insertion are not appropriate in all cases. The expression "selectivity of surface processes" is explicitly marked a non-variant of "process selection". It is, however, unclear what the underlying processes are.

Also "transmission electron microscope" is annotated a variant of "Transmission electron microscopy" in the test text and recovered as such from all four systems. While both compounds built on the same succession of lemmas, i.e. "transmit", "electron" and "microscope", they differ in their head words "microscope" vs. "microscopy". While the former is a thing, the latter a science. Whether such constructions are variants in all instances is doubtful.

It is certainly less doubt if morphological and/or derivational variation occurs in the non-head of the compound. For example, all four systems found "atom displacements" to be a variant of "atomic displacement". It is likely that this variation process will be much more reliable than derivational variation of the term's head word.

More in depth investigation is required to uncover similarities in the variation patterns and to examine the underlying mechanisms of applicability. It is also interesting to see whether the same variation mechanisms apply in an industrial application or whether some templates can be excluded.

One of the major drawbacks to more sophisticated term variation recognition in the English version of CLAT is due to the order of linguistic processing. In the current English CLAT architecture, terms are checked and matched without previous grammatical analysis of the text.

Experience in the German terminology checking has shown that noise can be considerably reduced when word analyses are disambiguated and syntactic tagging on a phrase level has taken place previous to term matching. Term recognition over phrase or, worse, sentence boundaries could thus be excluded.

6. Conclusion

In this paper we have presented CLAT, the Controlled Language Authoring Technology and in particular its terminology component. We compare the performance of CLAT's terminology tool in terms of precision and recall with two similar systems and discuss future development and research directions.

7. References

Carl, Michael, Johann Haller, Christoph Horschmann, Dieter Maas, and Jörg Schütz, 2002. The TETRIS Terminology Tool. *TAL, Structuration de terminologie,* 43(1).

Carl, Michael, Ecaterina Rascu, and Johann Haller, 2004. Using weighted abduction to align term variant translations in bilingual texts. In *Proceedings of LREC*.

Enguehard, Chantal, 2003. CoRRecT : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes. In *in Proceedings of TALN*.

Haller, Johann, Horschmann Christoph, Rita Nübel, Ursula Reuther, and Axel Theofilidis, 2002. Sprachtechnologie im Einsatz. Terminologie - Workflow - Sprachkontrolle in der Technischen Dokumentation. Technical report, IAI.

Jacquemin, Christian, 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Maas, Heinz-Dieter, 1995. Documentation of the features used in MPRO. Software documentation, IAI, Saarbrücken.

Maas, Heinz-Dieter, 1996. MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Roland Hausser (ed.), *Linguistische Verifikation, Sprache und Information*. Tübingen: Max Niemeyer Verlag.

[2] `http://www.sciences.univ-nantes.fr/info/perso/permanents/enguehard/`

RESEARCH IN TERMINOTICS IN CANADA AND QUEBEC
Current situation in 2004

Prof. Pierre Auger, director
Department of languages, linguistics and translation
Laval university
Quebec (Canada)
pierre.auger@lli.ulaval.ca

Abstract

Canada and Quebec, it is a well-known fact, constitute one of the cradles of the modern terminology in French language. This interest for terminology which goes back to the Sixties is not contradicted with the years. Well before the arrival of the information society, terminology had already joined data processing with the development of term banks built around terminological databases managed by databases management systems. One could also speak about the development of electronic dictionaries for the MT systems of first generation also called direct systems. Therefore its computerization from the very start of the Eighties was well initiated and immediately bonds were established with the research groups in automatic processing of natural language. Quebec and France were the first among the French-speaking countries to launch projects in automation of the terminographic process and development of software for assistance to the terminologists.

In the Nineties terminologists started at to name this new field as "terminoticss" which aims at the automation of repetitive terminographic tasks" and "the whole of the techniques and tools which contribute to automate the terminographic production chain. Some, to further go, created acronyms like MAHTER (Machine Assisted Human TERminology) and HAMTER (Human Assisted Machine Terminology).

1. History

In Canada, like elsewhere in Europe, the privileged bonds that maintained terminology with translation supported a "term to term"approach in terminology. It will be seen that several years later, under the influence of corpus linguistics, it is the textual approach which is privileged in terminology. It is in fact around these two poles that the history of automation in terminology is articulated.

The first period which goes from 1960 to 1980 approximately is one of stagnation in this sense that little progress was carried out after a first contact with automatic data processing. Should it be recalled that in the middle of *the Sixties*, whereas the automatic text processing was in its early stages, one already sees the interest to put the structured databases (DBMS and databases) technology at the service of terminology: association between file cards records / terminological data and information fields, association which is done naturally. Then one bumps constantly against difficulties then almost seen like insurmountable, as for example the display of the character sets of all " industrialized " languages (e.g. languages of East and Western Europe).

As for their data structure, term banks move little as it is often the case for terminography practices. Indeed, terminographic practices evolved little since standardization work on the contents of the terminological records (cf. Colloque International de Terminologie"Les données terminologiques (terminological data)" held in Baie St-Paul – Quebec in 1972). One observes a certain structura redundancy from a system to another resulting from traditional practices influenced by methods used in lexicography. Methodological works concerning terminographic practices, should it be pointed out, reached their apogee with Wüster and works from ISO Technical Committee ISO-TC37. The standardization of a unified terminological data model appears among these works. A period of stagnation then occurred touching term banks data models more centered on the description of term forms than on meaning more difficult to represent (> ontologies of years 2000): how to treat these more abstract entities known as concepts in terminology ? Progress was awaited on the side of documentary sciences to define universal classification plans like the UDC, but also of work from linguists in lexical semantics. It is only in the Nineties, that one will attend a certain release which will materialize in revision of the terminographic data model. The very synthetic data model that Sager provided makes it possible to appreciate the gaps between the data models in use in term banks and to identify the parts well described as the least well described as for the entities " conceptual specification ", " linguistic specification ", " pragmatic specification " and " FLequivalent specification " of Sager. There is a certain parallelism, source of hesitations between the lexical semantics model of the lexicologists and the notional models (pre-ontological) of the wüsterians.

Thus, the term banks grew bigger in number of terms available, without however undergoing software developments of importance. They remain primarily a business for computer specialists and continued to run on large computers what reduced considerably their portability.

During the second period which goes from 1980 to

1990, term banks increased their terminological diffusion thanks to the progress of telematics and personal computing. Moreover the phenomenon of " down-sizing " is getting stronger, this term refers to " the migration of systems from mainframe to minicomputers as wells as to desktop personal computers. The on-line access not yet supported by a nwetwork like Internet will take rise with servers and will make it possible to everyone to connect itself to term banks using pc's connected via the telephone line. The portability of term banks was thus the major stake of this decade, the difficulties of implementation of various nature (limits of capacity of the micro-computers of type PC's: less powerful processors, limited memory RAM, restrictive operating systems) for a technology like cd-rom to some extent masked the obsolete character of the term banks and their greater weaknesses like the inaptitude to represent the meaning (localization of the concepts in a given terminologicalfield) and the hierarchisation of various terminological uses as well (pragmatic of the lsp's, treatment of the terminological variation, hierarchical labelling of variants, usual collocations of terms, their syntagmatic derivation as well).

The third period which extends from about 1990 to 1995 directs itself towards a direct customer approach. The availability of new environments offered by Windows lets foresee a new user-friendliness in computers. The compulsory arrival of more standardized software made under Windows to make easier the use of the computers. Lastly, the new arrival of cd-roms will ensure the portabililty of term banks while waiting for the revolution brought by Internet. As major progress it is necessary to quote the new textual orientation of terminology that was made possible by new information technologies directed towards the text like the full-text databases managers using textual indexation. It is also during this same period that appeared the first works touching the automation of terminological work-chain, like machine analysis of texts and assisted term extraction. The data-processing tools offering assistance to terminography are various and make it possible to achieve the tasks, from the simplest (e.g. lexical information mining, statisticson words), to the most complex ones (e..g. intelligent extraction of, alignment in bi-texts, terminological and ontological extraction, assisted drafting of definitions).

In front of the general assesment that term banks remained, in spite of renewal of the interfaces, poor tools for conceptual representation of the term and its place in the broader onomasiologic universe of its field . In the beginning of the Nineties terminologists were searching for new means to improve the processing (or treatment) of the notional aspect of term. This orientation, in agreement with a wüsterian vision of terminology which grants the primacy of concept in the terminographic treatment (terminological lexicography according to Wüster term) brought new informational fields on terminology records like : - *hyperonyms,* mention of the higher immediate concept , - *hyponyms,* mention of the lower immediate concept , - *isonymes,* mention of a same level concept , etc.) in accordance to Wüster's proposals and in a way similar to the notional treatment carried out for the *Dictionnary of Machine-tools* without including explicit references to universal classification systems like UDC used by European documentalists at that time. One may find in the Seventies for standardization of terminological work by organizations like the ISO a comparable " systematic " approach. Another assesment may to be made here: the huge existing term banks did not arrive yet at this type of systematic representation because the lack of preliminary notional description work of terminological universes presented in these banks. It is difficult to imagine for the moment the use of automats for the management of the notional tree structures without preliminary human description. It is known in addition that the mini-term banks (unifield) manage better this aspect which does not any more appears as the titanic task of classification of the universe. The decade which follows will continue further the pairing *term and knowledge* with the rise of cognitive sciences and the computer assisted development of ontologies.

The fourth and last period which extends from 1995 to our days will see the arrival of the most significant progress with the development of " intelligent " automats for terminoticss able to proceed to the extraction of simple term as well as complex ones, of terminological variants, of defining segments as a tool for the assisted drafting of definitions. Finally a more semantic approach led the linguists engineers to develop automats for the extraction of ontologies.

The contemporary period which associates more and more the concept of terminology with the knowledge concept is the direct response to new orientations of automatic processing to support the expertise on the virtual world and recent progress of the software science and the knowledge engineering (artificial intelligence). Fundamental research came to assist the terminologists in their descriptions aiming the construction of automats being able in particular to extract ontologies starting from a corpus of specialized texts: how to locate the term describes in the specifically studied terminological universe? The other aspect which comes to enrich fundamental research in terminology and terminography as well is the treatment of denominative lexical variation in terminography. Here, the construction of automats for the localization of terminological variants constitutes the second pole of improvement of the term processing in terminography: how to re-use the terms described in the professional technical discourses?

By closing this point, it is advisable to precise that this short chronology corresponds well, but not exclusively, with the situation lived in Canada and Quebec.

2. Mixed approaches in terminoticss

In this technological complex, one may find *linguists – computing specialists* or *computing specialists – linguists* able to program automats with linguistic base resulting from fundamental research in NLP (Natural Language Processing). The language workers which are interested in the assistance for terminology which is to be associated to industrial research like for example the design of a working station for the terminologist. In 2004, one may find them all in the methodological consensus for corpus based terminography.

It is necessary also to look at other fields of research (we have already enumerated some of them) which nevertheless have direct relationship with terminoticss , here are some others:

- the use of Internet in terminography (sources, corpus, terminological watch)
- the intelligent information retrieval (data mining)
- the indexing and analysis of texts
- the extraction of ontologies
- the alignment of bi-texts
- translation memory
- autocorrection of texts
- the extraction of terms (CT ans ST)
- assistance with the drafting of definitions
- interfaces of input, frm text to DBMS
- terminology managing systems
- term banks
- co-operative telecommuting in wide-area networks
- the working station for the terminologist

3. The effective chaining of the various stages of terminographic work

Now let us examine the different stages which the contemporary terminologist must cross in the development of a terminological repertory of quality :

1. Preparation of the terminological work: the terminologist at this stage must achieve preliminary tasks which above all are of documentary nature (exploration of the field, search for documentation, selection and recording of the sources). This stage should normally be based on the use of Internet, a mean likely to provide the terminologist with basic knowledge on the speciality field he wants to explore, on the inventory also of reliable up to date documentary sources, in order to apprehend the overall structure of the field. The wide-area networks of information became impossible to circumvent at this stage as an initiation to a speciality field : consultation of specialized sites, technical ans scientific documents, bibliographies, catalogues of libraries etc. It is necessary to underline here that the step here exposed does not deny a more traditional approach, but includes it by means of digitalization of the printed texts.
Means used: the realization of this preliminary stage supposes on behalf of the terminologist a control of the navigation tools on the Web such as documentary robots (infobots), metasearchengines (e.g.Copernic, Metacrawler, All in One) and usual search engines " free " on Internet (AltaVista, Yahoo, Google, etc). There is also a crowd of search engines, more or less specialized, commercially available like Alceste, Lexibot, Discovery and many others.

2. Terminographical work as such: this stage refers to the achievement of terminographical tasks related to the processing general and specialized writings (formating and processing of a corpus of texts, establishment of the nomenclature, lexical analysis and extraction of terms (ST and CT), extraction of terminological and paraterminological data, drafting of definitions, pairing of equivalences between two languages etc).

The terminologist work, simplifying a little, consists to pile up and select texts (corpus) (1) from where he will extract valid data (2), to process (selecting and formatting) these data according to terminographical methods (3), then to store them generally in the form of databases (4) and finally to edit them to be diffused in the form of conventional terminological repertories (dictionaries, lexicons, vocabularies etc.) or in the form of electronic text (graphic, textual or hypertextuel) (5) (e.g. Auger: 1994). One can imagine right now a scenario of automation which will deal with each stage which has been just enumerated and which will ensure the chaining of each one of these stages. In fact, a scenario where starting from rough texts a complete terminological repertory is made up little by little and this, with a minimum of manual human interventions.
2.1. Preparation of the corpus: criteria for the selection and preparation of texts, (digitalization, edition, marking up, tagging): types, number, size of documents and corpus.
Means used: on line localisation of texts (with the help of robots, métaengines, search engines), downloading of electronic texts, preparation of digitalized texts using commercial text processors (like Word or WordPerfect, etc.), use of OCR softwares (e.g. Omnipage or TextBridge), use of documents presentation software (e.g. Adobe Acrobat).

2.2. Processing of the corpus: making of an indexed full-text database, extraction of ontologies, extraction of the ST and CT (nomenclature), of concordances (collocations,) extraction of valuable information for the assisted drafting of terminological definitions, assisted seek of variants (1 language) and equivalents (2 or more languages), extraction of ontologies.
Means used: use of textual analyzers (full-text indexing), of ontologies extractors (conceptual analysis), of automatic or semi-automatic term extractors, use of concordancers, automats to extract specific collocations, macros to search definitory segments for the assisted definition of terms, use of aligners with bilingual texts to locate equivalences. To mine in texts using commercial tools of full-text search (DTSearch, NatQuest, Isys, TACT, WordCruncher, WordSmith) using the technique of textual indexation which proved to be very useful and powerful tools for processing texts (evolution towards " intelligent " search or *data mining).* Simple concordancers (p. ex. MicroConcord, MicroOCP) appear among the oldest tools of assistance in lexicography and terminography. More recently " bilingual aligners ", new type of concordancers able to align a text and its translation (textual databases also called Bi-texts) developed for applications in TRAM (translation memory) which came to enrich the panoply of tools for assistance in terminography. This type of software allows, for example, to find all the contexts of a given request and their equivalents in the other language. TransSearch, WinAlign, Multiconcord appear among most known softwares of this type. Finally it seems necessary to speak about terminological extractors which occupies a center position in the terminographical process, a relatively recent software category where much development has been carried out these last years. This software makes it possible to extract all the occurrences of ST and CT of a text (or a corpus)

already digitalized. Among most known of this software for the extraction of terms, let us quote Adepte - Nomino, Lexter, System Quirk, TermFinder, Termplus Ztext. Adepte – Nomino is a software adapted to the processing of texts in), it gives often variable results for a same text, situation related closely to the type of algorithm used (according to whether a syntactic analysis is carried out or not). They have also the disadvantage to be generally attached to a language in particular, as it is the case in MT. Finally, there is a crowd of specific automats developed by university laboratories adapted to their own house-processing. In short, one can say that, until now, there are various terminographic scenarios of automation, that these scenarios are variously robust, but that none makes it possible to ensure the complete chaining of all the stages of terminological work. For the extraction of ontologies, a software like DockMan (I .Meyer, Lake Group, University of Ottawa) has been developed but it needs human assistance to give good results.

2.3. Storage and processing of data collected (design of databases and use of enriched, textual databases etc), making of relational terminological databases: DBMS / glossaries managing system, development in real time of virtual terminology records using peculiar automats or infobots.

Means used: use of DBMS for designing relational terminological database (e.g. Access, FileMaker, Oracle), of commercial glossaries managers (mini-termbanks managers: Multiterm, TermStar, TermBase, Termex). Developpement of automats for on line information searching capable of creating in real time virtual terminological records.

2.4. Physical storage of data: technical choice of data supports (cd-rom, optical discs, portable hard disks, flash memory etc), new perspectives offered by networking, the sharing of discs on distant host-computers.

Means used: material allowing the transfer of data (Zip drive, flash memories, optical disks, cd-rom. Dvd-rom etc), requirement of substantial knowledge concerning systems and materials from the terminologist.

2.5. Edition and diffusion of the end product: terminological dictionary accessible in various forms: portable products, products accessible on networks, products usable by other information processing systems, traditional printed products.

Means used: digitizing softwares and OCR system (Omnipage, TextBridge), documents indexers ans search engines (e.g.DTSearch, NatQuest), software for the displaying of documents (Acrobat Adobe), electronic forms of DTD type, desktop publishing software.

In short, all terminology projects comprise these 4 or 5 typical phases, but here stop the similarities, in reality, the differentiation of the needs and of the approaches (translational, real-terminological, "amenagiste", socioterminologic, textual discursive, cognitivist) made so that the ways of making for individuals and groups vary considerably and that it is illusory to think about the making of generic and universal terminotics tools. The automated working stations must necessarily take into account this constraint. The simplified model presented here is actually far from being simple in its application, the processes which must be implemented in the programming of the automats for assistance in terminography are extremely complex in accordance to human language, this explains the absence of acomplete functional terminotics model for the moment, even in a prototypic state.

However, in spite of these divergences, it is not useless to try to isolate the fundamental operations which implicates a systematic terminographic approach based on lsp text analysis, a particularly fertile current these last years resulting from corpus linguistics which makes possible the maximal use of the computer possibilities and of the softwares which governs them. In this approach shows through clearly the concept of workshop or terminological workstation where are led, around an automat for textual analysis (full-text analyzer), various strategies for extraction of terminological and paraterminologic informations, like the segmentation of text in its formal morphological and syntactic components (simple "words", uniterms, complex terms, collocations, phraseologisms, contexts), semantic and cognitive components (notional localisation and hierarchisation, synonymy and polysemia, definitory clauses, decoding of anaphoras) and pragmatic components (specific formulations of lsp texts, lexical stratification and terminological levels, terminological variation). These descriptions can moreover be led by the attribution of various properties to the formal items of the text (enrichment of the texts by various markings: grammatical, notional, sociolinguistic, pragmatic etc.) and quantified (use of terminometric statistics measurements: absolute and relative frequency, distribution, dispersion, etc).

In a second phase strategies for the selection of extracted information are conducted (e.g. directly extractable terminological data: entries , variants and synonyms, contexts, references (sources)), then those infered from the text : definitions, notes on the concept or its use). Indeed, the power of machine analysis and the exhaustiveness which it allows, produced, on the contrary of manual analysis, a proliferation of information which it is advisable to manage in a particular way to maximize the benefit for the terminologist. Such a strategy, for example, will make it possible to draw the best contexts from the corpus, to exploit the definitory clauses found in the texts for the drafting of definitions, to use the numerous contexts forsaken for the drafting of linguistic, encyclopaedic and notional notes. In short, to use the maximum of the information extracted following the various analysis and to use only this information. Taking into account the socioterminologic levels corresponding to the various types of specialized discourses (e.g. text typology and constitution of the corpus), many useful informations could be drawn from the analysis (e.g. simplification phenomenon, identification of professional slangs etc.).

In a third stage, all collected and sorted information is stored in the form of electronic database on an any support (text file, database, hypertext) to make the final terminological dictionary, accessible on line, in full-text or structured format, or in printed format. Future for these electronic dictionaries lies doubtless in the enriched full-text format, more dynamic than the structured format (traditional records made under DBMS).

Such a program does not go without posing many difficulties, the current computer stations for terminography remain primarily a collection of various softwares which run under an integrator of the Windows type, these softwares are not always compatible the ones with the others (routines for the transfer of information have to be written to facilitate the binding of each phase), they present - the textual analyzers at the first head - enormous difficulties for the average user, etc. In addition, the imposing arsenal of necessary software tools requires in priority the use of commercial tools available, rather than to have to develop expensive softwares and often less efficient. At the end, the greatest difficulty is to develop a very strict work protocol where all operations are well carried out, just as the integration of the bonds between the various stages in term of data transfer, in order to ensure an effective and safe information management.

3. Problems of automation in terminology

It is advisable to distinguish here two distinct aspects which require different solutions by distinct actors :

a) *The data container*

- Chaining of the terminographical work phases and general effectiveness of the process;
- Ruptures in the work chaining : need for data conversions, manual interventions, losses of information;
- Ergonomic aspects of the software system: software sometimes difficult to use, necessity for periodicals updating, additions of new tools, continuous training for the terminologist;
- Unilingual sophisticated tools designed to process only one language (parsers), a constraint which generates very significant costs;
- Availability of powerful tools for " intelligent terminological mining " under Internet (e.g. concept of " data mining ");
- Competition between developers which conducts to forget the terminologists needs;
- Availability of texts in electronic format (necessity to remedy at this situation).

b) The linguistic contents

- Problems touching linguistic analysis
  - Division of terms in ST and CT, phraseologisms;
  - manual desambiguisation necessary;
  - silence and noise produced by the automats;
- Treatment of the terminological variation
  - Identification of synonyms, morpho-lexical variants, attribution of the anaphoras;
  - Treatment of polysemia;
  - Assistance in definition of terms .
- Problems to represent the concepts and the treatment of ontologies

Many of these problems will not find a solution before several years whereas the computers will be able to manipulate the meaning in text processing and analysis.

4. Research touching terminotics in Canada-Quebec

Canada and Quebec in particular appear among the first French-speaking States to take interest in terminography and the update of its methods and their automation. Historically, they constitute the cradle of bilingual French-speaking terminology.

a) Government agencies

- *Office québécois de la langue française (Oqlf)*
  A pioneer among the main term banks with the *Grand Dictionnaire Terminologiqu*e (formerly known under the name of *Banque de terminologie du Québec* or *BTQ*), the organization, started in the Nineties a substantial reflexion on the topic of automation in terminology. The software platform Sami was designed to be used as an interface between the terminologists of the Office and the GDT, it is used through OQLF intranet and through Internet. There is also the software *Adepte – Nomino* which is a product resulting from this orientation, a collaboration between OQLF and UQÀM (Université du Québec à Montréal) its *Centre d'ATO*. The implication also of the RINT (Réseau international de néologie et de terminologie) made it possible to intensify this reflexion through the publication of the Journal *Terminologies nouvelles*, now known as the *Cahiers du RIFAL*. The same applies to its collaboration with the CCN CT37 ISO which also works in updating and modernization of working methods in terminography. The new orientation that took the OQLF a few years ago to base its terminology work only on the mining of information coming from Internet is resolutely to class among the terminotics practices.

- *Le Conseil du Trésor du Québec*
  This government agency which does not have anything to deal with linguistics a priori made work of pioneer in Quebec regarding document management in the middle of the Eighties. One finds here still the Centre d'ATO with Sato, its software for management and documentary indexing which permits to explore huge corpora of indexed texts and even to categorize all the words of a text to carry textual searching using grammatical patterns (e.g. N + prep + N). Terminologists early found interesting terminographic applications for *Sato*. Future has in fact confirmed the undeniable utility of this kind of software for terminology work.

- *Réseau international de néologie et de terminologie (RINT)*
  The network played during several years the role of project manager regarding information on Canadian works touching terminotics. Thus, it has setted up a continuous review of terminotics software, created the gate *Attrait* dedicated to terminotics on the Web site of the Rint and published in *Terminologies nouvelles* journal

many articles as a few issues on this theme .

- *Bureau de la Traduction (Gouvernement du Canada)*
  The *Bureau de la Traduction* is also well-known in the world thanks to its *Termium* term bank. Like its québécois counterpart, the Bureau focused early on the question of automation in its internal working methods, *Yvanhoe* software was developed a few years ago to be used as aninterface between the terminologists and Termium One also finds much information on the Web site of the Bureau. It is in close connection with the Canadian Observatory of the Language industries.

- *Otiaq (Ordre des traducteurs et interprètes agréés du Québec)*
  Otiaq (formerly *STQ – Société des traducteurs du Québec)*, historically, has always given interest in the working methods in use by terminologists and carried out his " virage technologique " a few years ago and followed-up its way considering automation as a priority for translators and terminologists. Moreover, some of its members are designers of software for terminotics. The professional organization organizes at various times of the year workshops to initiate its members to linguistic computing and data processing .

b) University laboratories

In a country as Canada where the translation is at the same time an activity socially very significant and remunerative, it is normal that university research is abundantly nourished with this source. At UQÀM, Centre d'ATO is well known mainly with its software *Sato* (a textual analyzer) and *Adepte-Nomino* (a term extractor for French language). RALI (Laboratory de recherche appliquée en linguistque informatique) at University of Montreal is especially known for its Bi-texts management tool known under the name *TransSearch* designed for intelligent mining in huge corpora, it also gave interest in the general field of linguistic automation. At the same university, Marie-Claude L'Homme and Patrick Drouin undertake research in terminotics (extraction of terms) with their group *Éclectik*, they have published many publications on the subject. The journal Meta edited at the same University has published many work relating to terminotics. At Université Laval, Jacques Ladouceur attached to Ciral works actively for the development of *TermPlus* software (a computerised workshop for terminotics). Lastly, let us mention the Lake Group at Université d'Ottawa directed by Ingrid Meyer, author of *DocKMan* software (formerly *Code* and *Ikarus)* directed towards the representation concept – term in the form of ontologies.

This too short panorama does not claim to be exhaustive, it does nothing but present the most significant work in the field.

c) Private companies

The place of choice which the documentary information retrieval occupies in software industry supported terminography in many aspects. Initially the terminological extraction (searching with keywords into documenticss) in corpora of texts occupies a significant place in the development of intelligent tools for textual mining, then the extraction of information of all natures in a corpus of texts coincide exactly with the working methods in use in terminography. Here are some achievements emanating of the data-processing sector.

- Alis Technologies (management of exotic alphabets)
- Ardilog: Natquest (textual search)
- Beetext (www.beetext.com (Bi-text)
- Copernic (metasearch engine)
- Documens (formerly Machina – Sapiens): automatic correctors, search engines, machine translation software
- Le Druide (idem)
- Gestar (Documentik and DocuStar, textual search)
- Terminologist (terminological management, Éd. De Lanaudière)
- MultiTrans and TermBase de MultiCorpora R & D Inc (terminological management)
- Sedrom-Sni (textual corpora)
- TermCruncher (terminological extractor)
- Terminotix (Logiterm, terminological extractor)
- Termis (terminological extractor, Printel)
- Traductix (Topograf and Atao, terminological extraction)

CONCLUSION

Terminotics is now a reality for human organizations and everywhere terminography is practised, every language professional must compose with that constraint. Considerable progress is carried out each year in this field thanks to the fundamental research and applied research in linguistic computing. Thus increasingly numerous commercial products for terminotics are put on the market each year it is the sign of a socio-professional implementation in rise. One can say that Canada and Quebec are not limping in regard with research and software products touching terminotics.

Valency Patterns of Danish Verbs as Terminological Knowledge Patterns

Lotte Weilgaard Christensen, Associate Professor
Department of Business Communication and Information Science
University of Southern Denmark, Engstien 1, DK-6000 Kolding
lotte@sitkom.sdu.dk

ABSTRACT

The aim of this article is to present a project dealing with the valency patterns of a number of Danish verbs as linguistic signals for the retrieval of terminological information from a Danish corpus on hydraulics. A valency theory called the Pronominal Approach will be implemented for this purpose. The verbs under scrutiny are first divided into three groups: metalinguistic term-related verbs, metalinguistic concept-related verbs, and relational verbs signalling hierarchical relations. I shall then deal with the questions of whether the valency patterns of the verbs studied are suitable as knowledge patterns and whether a relation of proportionality exists between their valency patterns and the categories of terminological information. Further, I shall discuss how and to what extent the Pronominal Approach offers a suitable and efficient search method for practical terminology work based on unstructured machine-readable corpora. The method presented will be tested and evaluated by means of a catalogue of knowledge patterns grouped according to the terminological information categories and applied to the machine-readable hydraulics corpus.

1 Background

This article describes some results of a project dealing with Danish verbs as linguistic signals. The aim of the method applied is to extract term-related as well as concept-related terminological information from unstructured machine-readable corpora. The need for such a method is particularly pressing because of the lack of adequate commercial language technology tools such as taggers and lemmatisers for Danish to support our search strategies. The method I employed was based on a study of whether (1) the valency patterns of the verbs under scrutiny are suitable as lexical knowledge patterns for the identification and retrieval of terminological information such as definitions and synonyms, of whether (2) the valency patterns, in their capacity of recurrent patterns, make it possible to create a more focussed approach than would have been possible using the bare forms of the verbs only, and of whether (3) a relation of proportionality exists between their valency patterns and the categories of terminological information. A valency theory called the Pronominal Approach (in the following called PA) was implemented for this purpose.

2 The Pronominal Approach

From 1993 to 1998 my colleagues and I participated in a research project called the UDOG project. The aim of the project was to explore the vocabulary and grammar of Danish. The primary objectives of our part of the project were to carry out a systematic valency description of Danish verbs and to compile a valency dictionary for both human and machine-readable purposes, primarily for machine translation.

The method we employed was a pronominal approach developed by a Belgian team and adapted to Danish by our colleagues in the project group (Daugaard & Kirchmeier-Andersen, 1995). The method of sense distinction applied by the PA is based on the pronominal substitution of the valency elements, the pronouns constituting a closed word class. The PA offers a reduced number of syntactico-semantic features which may be derived from the pronominal forms (Daugaard & Kirchmeier-Andersen, 1995: 4):

1. **Syntactic forms:** noun phrase, prepositional phrase, adverbial phrase, sentence (finite, non-finite)
2. **Syntactic functions:** subject, object, prepositional objects, valency-bound adverbials
3. **Semantic features:** human, concrete, abstract, countability, manner, direction, etc.

A basic assumption of the method is that a constant relation of proportionality exists between pronouns and their nominal constituents. According to the PA, (1) and (2) are changed into pronominal sentences as follows:

(1) *Teknikeren* adskiller *motoren*.
(Lit.: *The technician* disassembles *the engine*.)
→
Han adskiller *denne her*.
(Lit.: *He* disassembles *this one*.)

(2) *Pumpen* adskiller sig fra *motoren*.
(Lit.*: The pump* differs itself from *the engine*.)
→
Denne her adskiller sig fra *denne her*.
(Lit.: *This one* differs itself from *this one*.)

In (1) the pronoun *he* indicates that the subject has the semantic feature "human", and in (2) the pronoun *this one* indicates that the subject is "concrete". In (1) the object is a direct one, whereas in (2) it is a prepositional object, and both are "concrete". In both examples the noun *motor* (engine) is represented by the pronoun *denne her* (this one), which shows the proportionality between pronouns and nominal constituents. The verb *adskille* is used to illustrate my approach. It has five readings, two of which are suitable as terminological knowledge patterns (Weilgaard Christensen 2000). I shall show only three of its readings. In technical texts, the Danish verb *adskille* may have the meaning *disassemble* or the meaning *differ from*, the latter being the terminological reading.

The first reading corresponds to the pronominal sentence in (3):

(3) han adskiller denne her

(Lit.: he disassembles this one)

The sentence comprises the full valency pattern since the reading in question takes two obligatory arguments: a subject and a direct object.

According to the PA, the argument slots are defined from a surface syntactic point of view. Thus, we need to establish one entry for each syntactic variation of a verb, which is a very important characteristic of my approach. This may be seen from the two terminological readings, which are related ones. The second reading of the verb *adskille* is expressed by the pronominal sentence in (4):

(4) han/denne her/det adskiller sig fra ham/denne her/det
(Lit.: he/this one/it differs itself from him/this one/it)

This reading takes three obligatory arguments: a subject, a reflexive object *sig* (oneself), and a prepositional object with *fra* (from), the preposition selected by the verb. In this reading, the subject and the prepositional object indicate the presence of coordinate concepts in the text. The same applies to the third reading of the verb, shown in (5):

(5) han/denne her/det adskiller sig (fra ham/denne her/det)
ved denne her/det/at+infinitive
(Lit.: he/this one/it differ itself (from him/this one/it)
by this one/it/a Danish to+infinitive)

This reading takes three obligatory arguments and one optional argument, indicated by parentheses: a subject, a reflexive object *sig* (oneself), (a prepositional object with the selected preposition *fra* (from)), and an adverb with the selected preposition *ved* (by). *Han* (he), *denne her* (this one) and *det* (it) indicate that the syntactic arguments may be realised by the semantic features human, concrete or abstract. Provided that the optional argument is realised, this reading is partly identical with the second reading above and thus provides information about coordinate concepts. Moreover, the adverb with the preposition *ved* (by) gives information about how the subject and the prepositional object differ from each other, i.e. it indicates a distinguishing characteristic. Both terminological readings occur in the active voice only. The following example has been gleaned from the hydraulics corpus. In (6) we learn that in contrast to its coordinate concept *directional glide valve*, *directional seat valve* is characterised by being capable of shutting off without leakage.

(6) Retningssædeventilerne adskiller sig fra retnings-
gliderventiler ved at de kan afspærre uden lækage.
(Lit.: Directional seat valves differ themselves from
directional slide valves by their capability of shutting
off without leakage.)

3 Corpora

The empirical part of my study is based on two corpora, one a technical corpus within the domain of hydraulics and the other a popular science corpus. The valency patterns of the individual verbs have also been analysed on the basis of the readings of the verbs in question found in the Odense Valency Dictionary, which was a result of the dictionary part of the valency project mentioned above.

The hydraulics corpus has served three purposes: First, the verbs under scrutiny have been extracted from this corpus. Secondly, I have analysed the valency patterns of the verbs as they occur in this corpus. Thirdly, I have used the corpus for testing the applicability of my method for practical terminology purposes. Different genres are included in the hydraulics corpus, which consists at present of about 110,000 words.

Owing to certain shortcomings of the dictionary and to the limited size of the hydraulics corpus, in some cases I have had to consult another corpus systematically in order to achieve a full description of the relevant verbs. This corpus is called the Danish Dictionary Corpus (in the following called DDC), it consists of popular science texts and comprises 6 million words. The DDC has been lemmatised and thus permits search for specific word classes, whereas the hydraulics corpus consists of unstructured texts.

4 Delimitation of Verbs

From a terminological point of view, texts consist roughly of terms, other terminological elements, and the contexts surrounding those two groups of elements. My method focusses on the latter group to the extent that it is realised in the form of verbs.

My grouping of the verbs under scrutiny takes as its point of departure a presentation by Maurice Gross, cited at a conference in Copenhagen (1999). According to Gross, technical texts comprise three types of verbs: technical verbs (i.e. terms), metalinguistic verbs, and conjunctional verbs.

I divide the metalinguistic verbs into concept-related and term-related verbs, depending on whether they refer to the content or the expression side of the concepts. Concept-related verbs include such verbs as *definere (*define*)* and *karakterisere* (characterize), whereas term-related verbs include such verbs as *kalde* (call) and *betegne* (denote). The conjunctional verbs, which I choose to call relational verbs, are verbs signalling relations. This group includes such verbs as *bestå af* (consist of) and *tælle til* (belong to). I have dealt only with relational verbs describing generic and partitive hierarchical or coordinate relations. However, relational verbs might be subdivided into verbs describing various semantic relations, e.g. temporal and causal ones. I have investigated 25 bare forms of verbs, resulting in a total of 47 terminological readings.

5 The Suitability of the PA Valency Patterns as Terminological Knowledge Patterns

A comparison of the three groups of verbs showed that the results achieved for metalinguistic concept-related verbs and relational verbs are better than those achieved for metalinguistic termrelated verbs. Most Danish verbs are formed analytically by means of e.g. prepositional objects or particles. This fact and the fact that the PA defines the argument slots from a surface syntactic point of view mean that the valency patterns derived by this method are particularly suitable as search patterns.

Thus the superior results achieved for the first two groups of verbs may be attributed to the fact that their valency patterns tend to comprise prepositions which may be used as search patterns capturing specific terminological data. Also, a high degree of and in some cases even a constant relation of proportionality has been ascertained between the valency patterns of some of these verbs and the categories of terminological information. The coordinate concepts in connection with the verb *adskille* (differ from) illustrated this.

The inferior results achieved for metalinguistic term-related verbs should be attributed, among other things, to the fact that several of their readings take direct objects, which means that their valency patterns do not comprise prepositions to be applied for the search patterns. Besides, search results have shown a smaller degree of proportionality, i.e. a search for *kalde* (call) will result in hits containing different categories of terminological information. Moreover, the search results for term-related verbs comprise both term and concept related information whereas for the other two groups only concept-related information is captured.

An important result achieved through my study is that optional arguments which occur with prepositions have proved essential for the search results as the prepositions can help to eliminate terminologically uninteresting patterns (noise) and ensure a more focussed search. The same holds for free adverbs with a relatively consistent structure and frequency. In this way optional arguments and free adverbs with a consistent structure have become obligatory in my study and consequently also from a terminological point of view.

As a result, I shall argue that the character-string approach offered by the PA provides a suitable search method for precise identification of terminological information categories. The suitability of the approach applies especially to corpora consisting of unstructured texts without any kind of tags, which do not permit search for particular word classes.

6 Testing and Implementation of the Method Using a Catalogue of Knowledge Patterns

I have evaluated the suitability of my method for practical terminology by establishing and testing a catalogue of knowledge patterns based on the results of the valency analyses and subdivided according to the categories of terminological information searched for. The test was carried out on the hydraulics corpus.

In order to obtain an overview of the corpus in hand one should start by generating a frequency list. This list can be used to ascertain how many of the verbs under scrutiny are actually in the list and to mark them up. It should be stressed that these will be verb candidates only as the occurrence of a particular verb does not guarantee that any of its relevant terminological readings are actually present in the corpus.

The process should be divided into three stages. **Stage 1** consists in a preliminary ordering of the concepts of the domain in order to gain an overview of its conceptual apparatus. For the purpose of this preliminary ordering, a search for readings is performed according to the following sequence:

1. Readings identifying extensional definitions, e.g. *inddel* i* (divide into)*,* *opdel* i* (divide into)
2. Readings identifying broader and narrower concepts as well as coordinate concepts, i.e. parts of extensional definitions, e.g. *henregn* til* (include under)*, adskil* fra* (differ from)
3. Readings indicating criteria of subdivision, e.g. *definer* ud fra* (define by)*, inddel* efter* (divide by)*, skeln* på* (distinguish by)
4. Readings identifying definitions describing the parts of concepts, e.g. *bestå* af* (consist of)

This stage provides a good impression of the conceptual coverage of the corpus. The information gained in stage 1 is recorded for subsequent use in stage 2.

Provided that stage 1 has resulted in a sufficient amount of information, one may proceed to **stage 2**, in which one may begin to compose definitions.

5. Readings which may identify intensional definitions, e.g. *definer* som* (define as), *forstå* ved* (understand by)
6. Readings identifying common as well as distinguishing characteristics, e.g. *karakteriser* som* (characterise as), *gælde for* (apply to). This information should be compared to the results of 2.

In **stage 3**, the terminology is established by means of the term-related readings. The following sequence is suggested:

7. Readings which may indicate a criterion of designation, e.g. *benævn* efter* or *ud fra* (designate after or by), *betegn* efter* (denote after)
8. Readings which may identify synonyms or standardised terms, e.g. *benævn* som* (designate as), *betegn* som* (denote as)
9. Readings which may identify abbreviations etc., e.g. *betegn* med* or *ved* (denote by)

The test showed that a subdivision of the catalogue into strong and weak knowledge patterns is advisable, and it also showed that one should begin by searching on strong patterns with a high degree of or even a constant relation of proportionality between valency patterns and categories of terminological information.

10 of 11 readings of concept-related verbs may be characterised as strong, and 7 of those readings occur in the hydraulics corpus. 16 out of 20 readings of relational verbs are strong ones, and 13 of these readings occur in the hydraulics corpus. However, only 6 out of 16 readings of term-related verbs are strong, 4 of them found in the corpus.

I am convinced that a catalogue of the type I have proposed provides the terminologist with an efficient search method for practical terminology work. In cases in which the readings derived by the PA did not provide a suitable search pattern, it was quite often possible to find other linguistic signals. Thus, the catalogue should be expanded by the valency patterns of more verbs, their nominalizations, e.g. *benævnelse* (designation), and the valency patterns of those nominalizations, generic expressions denoting a hyponomy relation (generic-specific), such as *arter* (kinds), *typer* (types), *grupper* (groups), or paralinguistic patterns such as colon, sign of equation, and finally combinations of signals.

7 Concluding Remarks

Very detailed comparative LSP and LGP studies carried out within the valency project have confirmed that of all readings of a verb only a certain number will be realised in LSP texts. In addition, the evaluation of my method showed that the number of readings and thus the amount of noise eliminated through searches in the hydraulics corpus was in some cases limited as well.

Once again one may illustrate this point by means of the verb *adskille*, which in its terminological readings occurs only in the active voice, with a reflexive pronoun, and with the prepositions *fra* (from) and/or *ved* (by). The string *adskil** occurs 34 times in the hydraulics corpus. A search for coordinate concepts, using a combination of *adskil** and *fra* (differ & from), resulted in as little as 12 hits. But a search with the active form only resulted in 10 hits, just as a combined search for *adskiller* & *fra* resulted in the same number of hits. The conclusion is that the search with the active voice was fully capable of reducing noise. A similar search in the DDC on all verbal forms of *adskille* resulted in 392 hits. A search restricted to the active voice reduced the number of hits to 156, whereas a search on the active voice plus the reflexive pronoun *sig* (oneself) and the preposition *fra* (from) reduced the number of hits to 87.

This gives me reason to assume that my method is stronger in the case of corpora with a relatively large volume and a comparatively low level of abstraction and specialisation. In the case of corpora with a comparatively high level of specialisation such as the hydraulics corpus, however, its strength seems to lie in the fact that the results of the valency analysis (and the catalogue established on the basis of it) gives the terminologist useful knowledge of how even very precise searches may be performed for specific categories of terminological information using certain valency patterns, one example being the active voice of the verb *adskille* (differ).

The evaluation performed on the basis of the verbs analysed so far also showed that the highest level of precision could be reached in searches for concept-related information. This fact is particularly interesting since so far there has been a lack of methods for retrieving semantic relations and since language technology tools for extracting term-related information have been much easier to find.

The Danish National Research Council for the Humanities have stated, in a statement issued in connection with the announcement of a symposion on Danish language technology in May 2004: “Since Danish is spoken only by a small language community and since Denmark has too few large private enterprises capable of meeting the challenge, substantial public investment in Danish language technology is called for in the years to come.” One of the consequences of the lack of adequate commercial language technology tools is that as mentioned above machine-readable corpora in Danish are normally unstructured texts. However, I am convinced that the linguistic approach to terminological information retrieval by means of the PA alone is very useful in manual or rather intellectual terminology work based on machine-readable corpora in its present form. Nevertheless a vision for the future of terminological corpus work in Danish would include a complete catalogue of knowledge patterns based on the PA and integrated into a pattern based extraction tool in order to automate part of the terminological extraction and retrieval process.

8 References

Ahmad, K. & Rogers, M. (2001). Corpus Linguistics and Terminology Extraction. In G. Budin. & S. E. Wright (Eds.),Handbook of Terminology Management, Vol. 2, (pp. 725-760). John Benjamins B.V.

Daugaard, Jan & Kirchmeier-Andersen, Sabine (1995). The Odense Valency Dictionary Programme for Verb Coding. Jan Daugaard (Ed.), Odense Working Papers in Language and Communication No. 8 (pp. 3-35). Odense, Odense University

Statens Humanistiske Forskningsråd (2004). Discussion paper

Weilgaard Christensen, Lotte (2000). Danske verber som knowledge probes i terminologisk korpusarbejde. In Anita Nuopponen, Bertha Toft, & Johan Myking (Eds.), I terminologins tjänst, Festskrift för Heribert Picht på 60-årsdagen (pp. 243-279). Proceedings of the University of Vaasa, reports 59, Vaasa.

Semantic Relations between Concepts in Danish Domain Specific Texts

Lone Bo Sisseck

Dept. of Computational Linguistics
Copenhagen Business School
Bernhard Bangs Allé 17 B
DK-2000 Frederiksberg
Tel: +45 38 15 33 61
Fax: +45 38 15 38 20
lbs.id@cbs.dk

Abstract

This paper describes a fragment of an ongoing Ph.D. project that aims at identifying and extracting conceptual relations from Danish domain specific text. From a Danish corpus on nutrition, linguistic patterns indicating the generic relation have been extracted manually on the basis of an arbitrarily selected test corpus. These patterns, ranging over a hybrid of linguistic expressions, have then been tested on the whole corpus and the resulting data reveals an interesting point of departure for further studies. Out of 124 sentences where the linguistic marker actually is a candidate marker for a semantic relation between concepts, i.e. domain specific concepts are actually present in the sentence construction, 93 occurrences indicate a semantic relation between two or more concepts. The concepts have been mapped into a concept system and the result is compared with existing fragments of a manually build ontology.

1. Background

My project is related to the OntoQuery project[1], OQ, which aims at developing ontology based search methods. Since ontology based search systems require huge amounts of domain specific ontologies in order to be a success and since it is a very time consuming task to build an ontology manually it will be of great interest to find ways of automating at least part of the ontology generation. Earlier and ongoing projects throughout the world have already developed various methods of more or less automatic term and relation extraction from both a statistical and a linguistic approach. However, there have hardly been any attempts to investigate Danish texts in order to find linguistic patterns that could indicate relations between concepts.

2. The Generic Relation

In the (manually build) OQ ontology, which is based on the above-mentioned corpus on nutrition, only the generic *is-a* relation has been implemented in the prototype ontology so far. The nutrition ontology has been integrated with the SIMPLE-ontology[2]. The *is-a* relation is also known as the *generic relation* or the *type* relation.

The following definition of the generic relation is taken from the ISO standard on terminology work (ISO 1087-1): relation between two concepts where the intension of one of the concepts includes that of the other concept and at least one additional delimiting characteristic.

The relation between concepts described in the following will be illustrated by the expression *aR(b,…,n)* where a concept, *a,* is related to one or more concepts, b,…,*n*, by the relation *R*.

3. The Pilot Project

As a starting point I chose to initiate my investigation by solely looking for the generic relation as in OQ. I selected a few articles from the nutrition corpus, one about nutrition in general and one about Vitamin A. In these articles I located sets of two or more concepts that were related so that one of the concepts was a superordinate concept of the other(s). The linguistic patterns that indicated the relation are outlined out in Table 1.

The whole nutrition corpus has been investigated with the corpus tool WordSmith in order to find out if it is possible to locate semantic relations throughout the whole corpus by using the manually found linguistic markers as search patterns. Since the nutrition corpus is very small, around 20000 words, it was possible to go through every occurrence manually in order to decide whether we are dealing with an indicator of a semantic relation. The following conditions were set up:

- Primarily two or more domain specific concepts should be involved in the collocation pattern
- Secondly I should be able to recognize the presence of a semantic relation
- The relation should be the generic relation

It was then possible to eliminate many of the occurrences as semantic relations due to the fact that no domain specific concepts were involved. The remaining occurrences, i.e. the possible semantic relations, were then manually analyzed in order to see how many of them actually indicated the generic relation between two or more concepts. The results from this analysis are summarized in Table 2.

[1] The purpose of the interdisciplinary project OntoQuery is to develop theories and methods for content-based information retrieval. Partners in the project are the following Danish research institutions: Copenhagen Business School, the Danish Technical University, Roskilde University and the University of Copenhagen. The project is funded by the Danish Research Agency under the Information Technology Programme. For further information please see www.ontoquery.dk

[2] The Danish SIMPLE lexicon has been developed at the Centre for Language Technology in Denmark.

expression	aR(b,…,n) where R =	collocation example
at være (to be)	er (is)	Chrom **er** et sporstof (chromium **is** a tracer)
colon **:**	**:**	…vitaminer**:** B1, ,…,n (vitamins**:** B1,…,n)
parenthesis ( )	( )	…organisk jern **(**hæmoglobin, myoglobin**)** …(organic iron **(**haemoglobin, myoglobin**)**)
fx (e.g)	fx (e.g.)	…spormetaller, **fx** selen og kobber (…trace metals **e.g**. selenium and cobber)
omfatte (to include)	omfatter (include(s))	…næringsstoffer **omfatter** vitaminer og mineraler (…nutrients **include** vitamins and minerals)

Table 1: The *is-a* linguistic markers

expression	aR(b,…,n) where R =	oc-cren-ces #	pos-sible rela-tions #	act-tual gen - eric rela-tions #	act-tual gen - eric rela-tions %
at være (to be)	er (is)	447	65	45	70 %
colon **:**	**:**	70	37	29	78 %
parenthesis ( )	( )	11	9	8	88 %
fx (e.g)	fx (e.g.)	10	9	9	100 %
omfatte (to include)	omfatter (include(s))	4	4	2	50 %
Total			**124**	**93**	**75 %**

Table 2: Evaluation scheme

Note that apart from the 11 occurrences expressed by the parenthesis marker there where 106 occurrences with only one term inside the parenthesis. 84 of these related domain specific concepts e.g. *xerofthalmi (øjentørsot)(xerophthalmia (dry eyes)).* Following the above-mentioned conditions they indicate possible generic relations. However, 76 of the 84 (90%) are synonyms like in the example. The remaining 10% were not synonyms nor were they related concepts. So the 11 occurrences in the evaluation scheme all involve two or more concepts inside the parenthesis.

Out of 124 sentences where the linguistic marker actually is a candidate marker for a semantic relation between concepts, meaning that some domain specific concepts are actually present in the sentence construction, 93 occurrences indicate a semantic relation between two or more concepts, that is 75 %. One could argue that this is not an impressive result and it will be necessary to aim at an accuracy rate much higher. I must emphasize that this is an initial pilot project, and I feel highly motivated by the accuracy rate of 75 % to continue the investigation.

The next step will most certainly be to experiment by applying a statistical method to my training data in order to measure the effect. For that special purpose the naive Bayes classifier approach to Word Sense Disambiguation[3] will be considered as a working method. I have chosen this method because the linguistic markers, both signs and words, have different "senses" throughout the text. As I pointed out previously, the parenthesis can have the "generic" sense, the "synonym" sense and maybe also other senses. It would therefore be interesting to improve on the result by applying a statistical method. From my data I wish to estimate the probability of the *is-a* relational sense of each of the linguistic markers in combination with the probabilities of the presence of concepts in the surroundings.

4. The Ontology

The related concepts that have been extracted on the basis of the five linguistic markers can be seen in Table 3. It should be noted that the linguistic markers don't carry any proven information about the direction of the relation, and there has been no effort made so far to solve this super/subordinate concept problem.

A project at Syddansk Universitet (Weilgaard, 2000) has investigated the suitability of a number of Danish verbs as knowledge probes for the retrieval of definitions in a Danish corpus on Hydraulics. One of the advantages of this method is that when using a definitorial verb as search pattern, meaning a verb that carries some kind of concept definition, a picture emerges of the relational patterns between the involved concepts. That means that the linguistic marker also indicates which concept is superordinate to the other and vice versa.

Despite the super/subordinate concept problem, the concepts have been manually mapped into a concept system. The decision about which concept should be the superordinate concept was based on intuition. In the collocation example from table 1, *cromium is a tracer*, the concept *tracer* is the superordinate concept because intuitively there can be different kinds of *tracers*. In the collocation example, *vitamins (B1,…,n)*, the concept *vitamins* is the superordinate concept because intuitively there are several kinds of *vitamins*.

In order to see if there is any reason in trying to relate concepts solely on the basis of the retrieved linguistic markers, the ontology is compared to the similar fragments from the manually build ontology in the OntoQuery project.

The Ontology built on the basis of the concept relationships Extracted from the Linguistic Markers, referred to as OELM in the following, can be seen in Figure 1.

[3] The naïve Bayes classifier approach to WSD is based on the premise that choosing the best sense for an input vector amounts to choosing the most probable sense given that vector (Jurafsky & Martin 2000, p. 638).

ernæringsproblem, *nutrition problem*	sult, *hunger* underernæring, *malnutrition*
ernæringsproblem, *nutrition problem*	overvægt, *overweight* undervægt, *underweight* psykiske spiseforstyrrelser, *psyk. eating disorder*
ernæringsproblem, *nutrition problem*	A-vitaminmangel*, vitamin A deficiency*
næringsstof*, nutrition substance*	zink, *zinc*
næringsstof*, nutrition substance*	riboflavin, *riboflavin* selen, *selenium* zink, *zinc* kobber, *copper* mangan, *manganese*
næringsstof*, nutrition substance*	fedt, *fat* kulhydrat, *carpohydrate* protein, *protein*
næringsstof*, nutrition substance*	vitamin, *vitamin* mineral, *mineral*
næringsstof*, nutrition substance*	kulhydrat, *carpohydrate* fedt, *fat* protein, *protein*
næringsstof, essentiel, *essentiel nutrition substance*	natrium, *sodium* klor, *chlorine*
næringsstof, essentiel, *essentiel nutrition substance*	vitamin, *vitamin* mineral, *mineral* sporstoffer, *tracer*
signalstof	nitrogenoxid, *nitrogen*
signalstof	serotonin, *serotonin*
signalstof	adrenalin, serotonin, kreatin
sporstof*, tracer*	chrom, *chromium*
sporstof*, tracer*	jern, *iron* zink, *zinc*
sporstof, essentiel*, essentiel tracer*	kobber, *copper* molybdæn, *molydenum*
stof, *substance*	steroid, *steroid*
stof, *substance*	selen, *selenium*
stof, *substance*	bilirubin, *bilirubin* ubiguinon, *ubiguinon* glutathione, *glutathion*
stof, livsvigtigt, *vital substance*	magnesium
vitamin, *vitamin*	A, B1, B2, B6, B12, C, D, E,
vitamin, fedtopløselig, *fat soluble vitamin*	A, D, E, K

Table 3: related concepts

When I extracted a similar fragment of the manually built OntoQuery ontology, an ontology emerged that showed a certain degree of similarity but it also contains substantial differences. The OntoQuery Ontology Fragment, referred to as OOF in the following, can be seen in Figure 2.

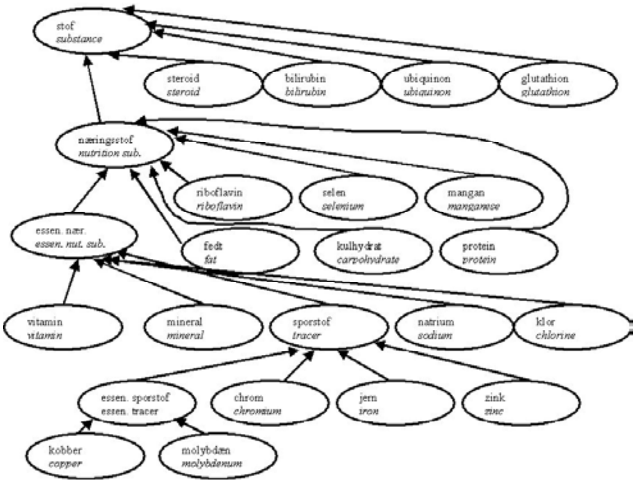


Figure 1: Ontology Extracted from Linguistic Markers, OELM

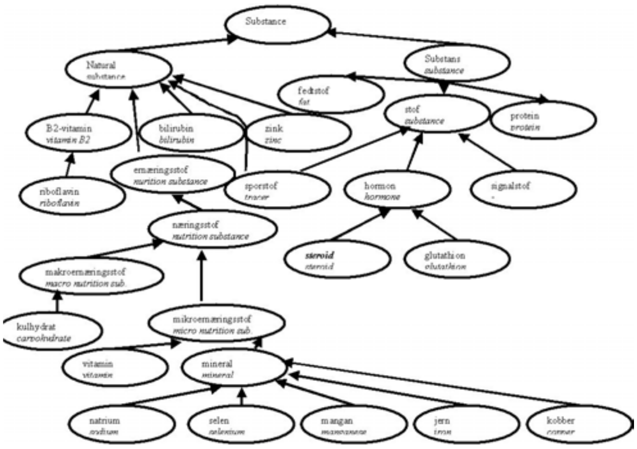


Figure 2: The OntoQuery Ontology Fragment, OOF.

For example, in the OOF there is a "*hormon*"-level in between the superordinate concept *substance* and the two subordinate concepts *steroid* and *glutahion*, whereas the OELM tends to be more general, e.g. *steroid* and *glutahion* are directly subordinate concepts to *substance* on the same level as *bilirubin* and *ubiquinon*, not indicating that *glutahion* and *bilirubin* are *hormones*. Some of these details will maybe be captured in an OELM that covers a greater amount of linguistic patterns and more data.

In general the OOF seems to contain more levels. Since it has been integrated with the SIMPLE-ontology, many of the concepts have been attached to already existing concepts. This means that some of the concepts have been attached to more detailed superordinate concepts or even more general ones depending on the concept in question.

However, the resulting concept relations in the OELM seem to be adequate for an initial ontology. When terminologists work with domain modeling the result will always depend on means and purpose. In the present example, a nutrition expert should verify if the OELM is an adequate model or not. Surely an expert will complain and the ontology must then undergo corrections. Not only is the ontology building process a very time consuming task in itself, it is also an iterative and dynamic process because it requires making corrections when discovering new terminological information.

5. Conclusion and future work

In this paper I have presented the initial results of a Ph.D. project that aims at identifying semantic relations between concepts in Danish domain specific texts. Based on a small set of data, it has been possible to construct an initial ontology based on the information that was carried by the identified linguistic markers.

The next step will most certainly be to experiment by applying a statistical method to my training data in order to measure the effect. For that special purpose the naive Bayes classifier approach to Word Sense Disambiguation will be considered as a working method.

It is interesting to find linguistic markers in order to state semantic relations between concepts, and implemented as a linguistic tool, it could save terminologists a lot of time. One of my hypotheses is, however, that the linguistic patterns that indicate semantic relations could vary from domain to domain. Therefore the linguistic markers are to be tested in a corpus from a different domain. In order to test this hypothesis I assume that I will need a larger collection of linguistic markers, e.g. more verbs, as the markers presented in this paper are not very likely to vary across domains.

Another very interesting experiment would be to merge an automatic relation extraction module with the terminological tool for construction of concept systems, within the framework of the project CAOS (Madsen et al., 2002). Since the aim of CAOS is to aid the terminologist in organizing and controlling the concepts and their relational features, this merging would probably facilitate the process even more.

6. References

Jurafsky D. & Martin J.H. 2000: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall, 2000.

Madsen, Bodil Nistrup, Thomsen, Hanne Erdman & Vikner, Carl: Computer Assisted Ontology Structuring. In: *Melby, Alan (ed.): Proceedings of TKE '02 - Terminology and Knowledge Engineering, INRIA, Nancy, 2002, s. 77-82.*

TERMINOLOGY WORK — VOCABULARY — Part 1: Theory and application
(Partial revision of ISO 1087:1990)

Weilgaard, Lotte 2002. ”Danish Verbs as Knowledge Probes in Corpus-based Terminology Work”. In: *LSP & Professional Communication, Volume 2, Number 2, October pp. 77-93.*

WordSmith http://www.liv.ac.uk/~ms2928/

CAOS – a support system for ontology structuring

Bodil Nistrup Madsen, Hanne Erdman Thomsen, Carl Vikner

Center for Terminological Ontologies,
Department of Computational Linguistics, Copenhagen Business School,
Bernhard Bangs Allé 17B, DK-2000 Frederiksberg, Denmark
{bnm, het, cv}.id@cbs.dk

Abstract

In the paper we describe a prototype system for terminological ontology structuring. The goal is to formalize traditional concept systems, in order to enable semi-automatic construction of concept systems, or ontologies. Feature structures are used as the formal framework, allowing for automatic treatment of the inheritance of characteristics, and various inferences concerning characteristics. These inferences are to be utilized to develop methods for supporting users with a terminological background in the construction of ontologies. In the paper we will focus on some of the methods developed so far, and explain their background and their functionality in the prototype system. Some of these methods concern the structuring of the ontology, i.e. how to support the user in maintaining consistency in the use of feature specifications. Other methods, not yet implemented, are devoted to support in identifying criteria of subdivision, which are crucial for definition writing in a later stage of the terminological working process

1. Introduction

In terminology work, concept systems are an indispensable tool for bringing order in the chaos of terms that translators and others often meet when they are going to work on texts belonging to a particular technical domain. However, it is a very difficult and time-consuming job to build one's own concept system for a particular domain, and a lot of time and effort is spend on the task of correcting drafts, even when the system has grown only minimally complicated.

Therefore the Department of Computational Linguistics and the DANTERMcenter are working on a project whose aim it is to develop a computer system designed to facilitate this job: CAOS - Computer-Aided Ontology Structuring (Madsen et al., 1999). A necessary prerequisite for this system development is a formalization of terminology work itself, as this has normally been carried out in a rather informal manner. Thus, in the project CAOS, the informal concept systems are replaced by formal ontologies.

2. Goal of the paper

In this paper we want to demonstrate how the use of formal feature structures in terminological ontology building can aid terminologists in the construction of concept systems.

The paper is organized as follows. In Section 3 we discuss some related projects. In the following section we describe how concept systems are formalized by means of formal feature structures for modelling inheritance of characteristics. In Section 5 we show how feature specifications are used in the CAOS application for ontology structuring. Finally, in Section 6, we illustrate how subdivision criteria of traditional terminology work are to be handled formally in our system.

3. Related work

Another project concerned with knowledge-based methods for terminology work is the COGNITERM Project[1]. In this project, a formalization akin to feature structures is used to represent characteristics of concepts, and the systems developed include automatic inheritance, but the consistency checks that we introduce in our system go beyond the detection of value-clashes described in (Meyer et al. 1997: 116).

Cimino (2001) describes methods for using knowledge to automatically or interactively position concepts in a hierarchy with multiple inheritance. The approach is similar to ours, but the system (MED) is designed specifically for the medical area and operates with a closed set of attributes (relationships + attributes in Cimino's terminology), where we suggest an open set, because we have found that in specialized fields, the attributes of delimiting characteristics may be very specialized and thus cannot be chosen before terminology work starts.

4. Formalizing terminological concept systems

The information about concept systems currently registered in terminological databases in general is not formalized, and it has to be worked out by the terminologist prior to entering data into the database, typically with no use of formal tools. To illustrate the informality of the data, let us mention that in the term base application DANTERM^{CBS} (where CBS stands for Copenhagen Business School), information about a related concept is represented by a term-expression denoting the concept, rather than by a concept-related information category, such as an ID-number which – in contrast to terms – uniquely identifies the concepts. Registration of the ID-number in the database would still allow the interface to present the user with one or more of the (synonymous) term-expressions.

Recently we have suggested that terminologists should use formal feature specifications[2] (consisting of an attribute and an associated value) to model the characteristics of concepts (see Thomsen 1998, 1999 and Madsen 1998b). This is similar to the approach described in Meyer et al. (1997). In our formalization, we operate with monotonic multiple inheritance. In the current version we have no

[1] http://aix1.uottawa.ca/~imeyer/research.htm

[2] Formal feature specifications, see Carpenter (1992)

hierarchy of values, but we are considering this for later versions.

As an illustration of our approach, consider the concept system in Figure 1:

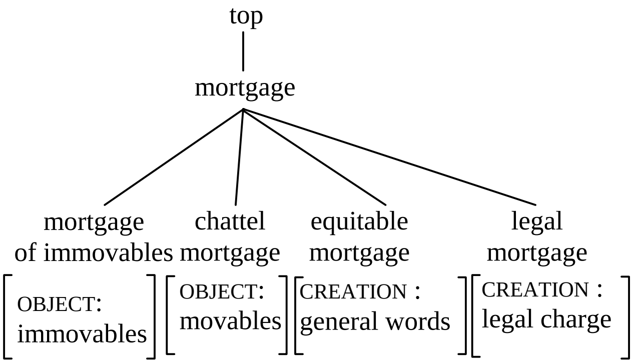


Figure 1: Concept system with feature specifications.

The concept *equitable mortgage* has the characteristic 'created by general words' (herein differing from a legal mortgage), which is found in a written source. This can be represented, e.g., by ascribing the feature specification [CREATION: *general words*] to the concept, as illustrated in Figure 1. Now, all concepts added as subordinate concepts of *equitable mortgage* will share this characteristic with it. This is reflected formally by the inheritance of the relevant feature specification to all descendants of *equitable mortgage*.

The formalization allows the development of computer systems that can carry out consistency checks, as described in the following sections. Another advantage of the formalization is that data becomes accessible for other purposes where formal ontologies are required, such as e.g. ontology-based information retrieval.

5. Integration of feature specifications into the ontology

5.1. Inheritance

In our system, a feature specification may be associated with a concept in two ways. It may be assigned directly to the concept, in which case we shall call it a primary (occurrence of a) feature specification, or it may be inherited from a superordinate concept. A feature specification is primary if it is the topmost occurrence in the ontology of the specification in question, all other occurrences of that specification being inherited. That means that a feature specification may only occur once in the ontology as primary. Primary feature specifications are registered in a table in the extended database. Inherited feature specifications, on the other hand, are not listed explicitly in the database but are computed on the basis of the ontology structure when needed.

Primary feature specifications may only be added to the ontology as a result of direct user intervention. Attributes and values are chosen from pick-lists or, in the case of new attributes or values, entered by the user. The attribute pick-list by default shows all the attributes used in the current concept system, and the value pick-list is limited to the values used with the current attribute in the current concept system.

Whenever a user wants to add a primary feature specification, the system carries out a number of checks. For instance, the ontology is tested for other occurrences of the same feature specification. If there is such an occurrence, the system warns the user, and we are developing methods for the system to propose different actions to be taken. If the other occurrence is on a subconcept of the current concept, the system should ask the user if she wants the primary occurrence to be "lifted" to the current concept. If the other occurrence is on a concept that bears no direct subordination relation to the current concept, the system should check whether it is possible to establish a polyhierarchical relation, cf. below in 5.3.

5.2. Insertion of new concepts

The user may ask for the insertion of a concept into a particular position of the ontology. The user must indicate this position by identifying the landing mother, i.e. the concept in the ontology that is to be the nearest superordinate concept of the new concept.

If the new concept shall be inserted into an intermediate position, in the current version, the user will have to move those concepts among the original daughters of the landing mother that are to be subconcepts of the new concept and check for inheritance inconsistencies manually.

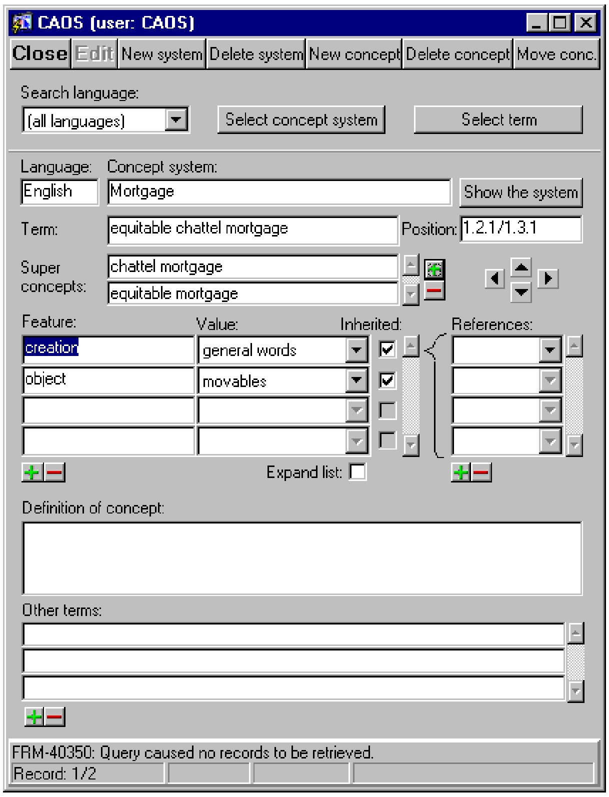
All the feature specifications of the landing mother will be inherited by the new concept. Then the user may add primary feature specifications to be associated with the new concept. Each feature specification indicated by the user will be dealt with as an addition of a new feature specification as mentioned above, and appropriate checks will be carried out. In future versions, these checks are to apply also to movement of concepts, cf. above.

5.3. Establishing a polyhierarchy

Suppose that the user wants to insert the concept *equitable chattel mortgage* into the ontology shown in Figures 1. Presuming that the user knows from her sources that *equitable chattel mortgage* is a type of *chattel mortgage*, and further that it is created by general words, she will want to insert *equitable chattel mortgage* as a daughter of the concept *chattel mortgage*. When this is done, the system tells the user, that *equitable chattel mortgage* will inherit the feature specification [OBJECT: *movables*] from its landing mother.

The user will accept this, and when prompted for primary feature specifications of *equitable chattel mortgage* she will give [CREATION: *general words*]. Now, while performing the checks mentioned above, the system finds out that this feature specification is already present on the concept *equitable mortgage*. Therefore it will warn the user that the addition is impossible as the feature specification occurs on *equitable mortgage*. In the current version, the user will have to figure out for herself that *equitable chattel mortgage* could inherit the feature specification in question from the concept *equitable mortgage*. This can be done by establishing a polyhierarchy with *equitable chattel mortgage* having both *chattel mortgage* and *equitable mortgage* as superordinate concepts.

The polyhierarchy is established by adding one more super concept to *equitable chattel mortgage* as illustrated in Figure 2[3]. Note that the position notation, 1.2.1/1.3.1, is calculated automatically by the system. Procedures to allow the system to propose this polyhierarchy are planned, but already giving the user information on where a given feature specification is used, helps the user correct her concept system.



Search language: (all languages)

Language: English

Concept system: Mortgage

Term: equitable chattel mortgage

Position: 1.2.1/1.3.1

Super concepts: chattel mortgage; equitable mortgage

Feature:	Value:	Inherited:	References:
creation	general words	☑	
object	movables	☑	
		□	
		□	

Expand list: □

FRM-40350: Query caused no records to be retrieved.

Record: 1/2

Figure 2: Multiple inheritance in CAOS

6. Subdivision criteria

In terminology work, subdivision criteria are used to group sister concepts in a concept system according to the features that characterize them, and in terminography such a criterion should be used in the *differentia specifica* of the definition.

In CAOS, the plan is to handle subdivision criteria via so-called dimensions and dimension specifications. A dimension of a concept is an attribute whose possible values allow a distinction between some of the subconcepts of the concept in question. Consider, for instance, the concept *mortgage* in Figure 1. Here the values *general words* and *legal charge* of the attribute CREATION make it possible to distinguish between two of the subconcepts, i.e. *equitable mortgage* and *legal mortgage*. Consequently, CREATION is a dimension of the concept *mortgage*. A dimension specification is the association of a dimension with its possible values. Thus, (CREATION: [general words, legal charge]) is a dimension specification on the concept *mortgage*.

There is an obvious interdependence between, on the one hand, dimension specifications on a concept and, on the other hand, primary feature specifications on its nearest subconcepts. For instance, if the concept *mortgage* has the dimension specification (CREATION: [general words, legal charge]), then two of its nearest subconcepts must have the primary feature specifications [CREATION: general words] and [CREATION: legal charge], respectively. Furthermore, it can be seen from the representation in Figure 1 that *mortgage* has two dimensions, i.e. OBJECT and CREATION, each distinguishing a separate set of subconcepts.

The dimension specifications of every concept are registered in the database. Whenever a primary feature specification is added to a concept, the corresponding pieces of information are registered in a dimension specification on the nearest superordinate concept. In future prototypes, it will be possible for the user to add information about dimension specifications directly on the superordinate concept, e.g. when a source mentions that "there are various types of mortgage, classified according to the type of the object mortgaged". In such cases, CAOS shall see to it that the corresponding primary feature specifications are added (interactively) to the relevant daughter concepts, possibly creating new daughter concepts if needed.

It may seem redundant to register this information both as a dimension specification on the superconcept and as feature specifications on the subconcepts. Our reason for doing this is that we also plan to register the source of the information. At a later stage of the work it may be relevant for the terminologist to know whether a dimension has been introduced because it was mentioned as such in a source, or because it was found as an attribute on two or more subconcepts. In the first case, moving the subconcepts to another position in the concept system shall imply that they 'lose' the attribute in question, whereas in the last case such a movement shall cause the dimension specification to be moved as well.

In some cases, sister concepts may be distinguished by more than one dimension, and in order to write proper definitions the user must make a choice between them. We use the term "subdividing dimension" to refer to that distinguished dimension to be used in the definitions of those immediate subconcepts where the subdividing dimension is included in the feature structure. Hence, the subdividing dimension is a formalization of the subdivision criterion. The promotion of dimensions to the rank of subdividing dimension is to be carried out in a final normalization phase not yet implemented, where the system is to present the user with a list of the relevant dimensions and the corresponding concepts, from which the user chooses the dimension to promote.

In all those cases where a set of sister concepts are distinguished by means of only one dimension, this dimension should automatically be promoted to subdividing dimension.

In the *mortgage* example, for instance, it is planned that CAOS should be able to find out that there are two groups of sister concepts, each carrying only one dimension. Accordingly, CAOS shall propose that the dimensions

[3] In the current prototype, consistency checks are not performed when establishing a polyhierarchy.

OBJECT and CREATION both be declared subdividing dimensions, thus creating two subdivision groups of subconcepts.

In Figure 3 we show an example with a graphic display of the two dimension specifications on the concept *mortgage* and the corresponding subdividing dimensions OBJECT and CREATION displayed in boxes covering the branches leading to the subconcepts concerned. As can be seen, the subdividing dimensions are very helpful to the user, as they help give a clearer overview of the field. Therefore we find it very important to incorporate subdivision criteria in a terminology management system, and to our knowledge, no other (semi-)automatic system does this.

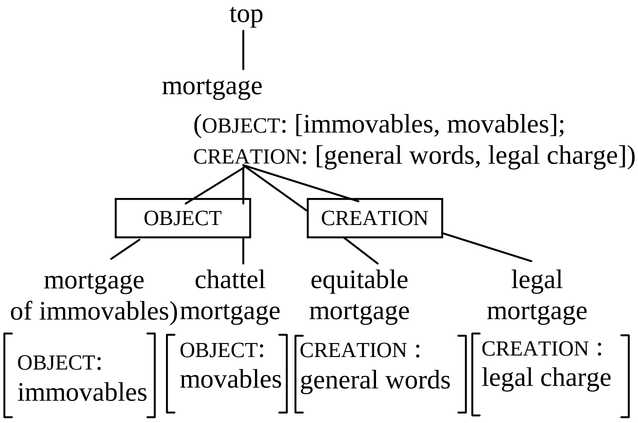


Figure 3: Subdividing dimensions and dimension specifications

7. Concluding remarks

We have presented our work on a terminology database which includes formalized information on concept systems, or ontologies, and we have shown how this formalized knowledge is then utilized to aid terminologists in their construction of concept systems by means of an automatic treatment of the inheritance of characteristics.

Future research includes development of methods for making more sophisticated feed-back to users in the case of feature and value clashes.

8. Acknowledgements

The project *Computer Aided Ontology Structuring* has been supported by the Danish Research Councils. Database programming in CAOS has been done by Bo Krantz Simonsen.

9. References

Carpenter, Bob, 1992. The Logic of Typed Feature Structures. Cambridge, Mass.: Cambridge University Press

Cimino, James. J. 2001. “Knowledge-based Terminology Managament in Medicine” In: D. Bourigault, C. Jacquemin and M.-C. L’Homme (eds.). 2001. *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins: 111-126.

DIN 2331: Begriffssysteme und ihre Darstellung. Apr.1980, Deutsches Institut für Normung. Berlin: Beuth Verlag GmbH.

Hull, Anthony; Bodil Nistrup Madsen & Hanne Erdman Thomsen, 1998. "DANTERMCBS for Everyone". In: *TAMA '98 – Terminology in Advanced Microcomputer Applications. Proceedings of the 4th TermNet Symposium: Tools for Multilingual Communication*. Wien: TermNet: 67-85.

Jacquemin, Christian, 2001. Spotting and Discovering Terms through Natural Language Processing. Cambridge, Mass./London: The MIT Press.

Madsen, Bodil Nistrup, 1998a. "The DANTERM Concept". In: *TAMA '98 – Terminology in Advanced Microcomputer Applications. Proceedings of the 4th TermNet Symposium: Tools for Multilingual Communication*. Wien: TermNet: 67-85.

Madsen, Bodil Nistrup. 1998b. Typed Feature Structures for Terminology Work - Part I. In: *LSP - Identity and Interface - Research, Knowledge and Society. Proceedings of the 11th European Symposium on Language for Special Purposes*. Copenhagen, August 1997, Copenhagen Business School: 339-348.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen and Carl Vikner, 1999. The project "Computer-Aided Ontology Structuring"(CAOS). In: *World Knowledge and Natural Language Analysis. Copenhagen Studies of Language* vol.23, Copenhagen: Samfundslitteratur: 9-38.

Meyer, Ingrid, Karen Eck and Douglas Skuce. 1997. “Systematic Concept Analysis within a Knowledge-Based Approach to Terminology”. In: S.E.Wright and G.Budin (eds.). 1997. *Handbook of Terminology Management*. Amsterdam/Philadelphia: John Benjamins.

Thomsen, Hanne Erdman. 1998. Typed Feature Structures for Terminology Work - Part II. In: *LSP - Identity and Interface - Research, Knowledge and Society. Proceedings of the 11th European Symposium on Language for Special Purposes*. Copenhagen, August 1997, Copenhagen Business School, 349-359.

Thomsen, Hanne Erdman, 1999. 'Typed Feature Specifications for establishing Terminological Equivalence Relations'. In: World Knowledge and Natural Language Analysis. Copenhagen Studies of Language, vol.23, Copenhagen: Samfundslitteratur: 39-55.

The GENOMA-KB project: a concept based term enlargement system

Judit Feliu, John Jairo Giraldo, Vanesa Vidal, Jorge Vivaldi, M. Teresa Cabré

Institute for Applied Linguistics
La Rambla, 30-32; 08002 Barcelona, Spain
{judit.feliu; john.giraldo; vanesa.vidal; jorge.vivaldi; teresa.cabre}@upf.edu

Abstract

The GENOMA-KB knowledge base includes four independent modules: a textual database, a factual database, a terminological database and an ontology. We will briefly introduce in this paper the main features concerning each one of the modules, and we will highlight the process of enlarging both the term base and the ontology.

Introduction

In the framework of the GENOMA-KB project, a special relevance has been given to the interaction between the ontology and the terminological information organized into modules. The first goal of this paper is to describe the process of enlarging and updating the term base module both for the linguistic information and the terms' direct link to the ontology. The second aim is to propose a combination methods strategy oriented to retrieve terms and term candidates for their inclusion in the term database. Some final conclusions and future research lines will be drawn at the end of this paper. A detailed discussion about the GENOMA-KB project is presented in a separated paper.

The GENOMA-KB description

In the genetics domain, there are some databanks publicly available, such as, Gene Ontology[1], LocusLink[2], Gene Ontology Browser[3], and GeneCards[4]. These resources include high-specialized data and they are an important assistance for domain researchers. However, their exploitation is difficult for other kind of users as terminologists, translators and scientific journalists.

From the terminology point of view, it is worth mentioning the attempts for the integration between terminological units and some conceptual information found at (Meyer et al., 1992; Condamines et al., 2000).

The GENOMA-KB integrates four modules: a textual database that contains specialised texts of this particular domain; a factographic and documental database containing the metainformation about the tagged texts in the corpus; a terminological database including the linguistic units transferring specialized knowledge, and a human genome ontology, which will be the basis for establishing a conceptual link between terminological units and the concepts they transfer. Figure 1 shows the tight relation between the four modules that take part in this knowledge base:

- Textual database: it contains actual documents directly related to the human genome domain. We collect texts in three languages: Catalan, Spanish and English.
- Document and factographic database: it registers bibliographic information about the texts in the textual database and metadata related to the genome domain.
- Terminological database: specialized knowledge units extracted from texts are introduced in this database and they are linked to concepts in the ontology.
- Ontology: concepts and their corresponding knowledge units entered at the terminological database appear in a knowledge organization based on a set of both hierarchical and non-hierarchical conceptual relations.

The term base and the ontology modules are closely tied due to the theoretical approach and its derived design decision of the software application to be used for managing the terminological and the conceptual information. The term base and the ontology editors are part of the OntoTerm[5] terminological management system. One of the application requirements is the previous development of the ontology, which will be the basis for the inclusion of any new term. This obligatory assignment has some theoretical consequences such as biunivocity between concepts and terms. One of our concerns has been to overcome this restriction by enlarging the number of conceptual relations and allowing multiple inheritance.

The ontology is built upon a initial set of concepts that have to be used as the starting point. These top concepts mainly concern the basic semantic categories, that is, events, objects, relations and properties. At present, the Human Genome ontology gathers more than five hundred concepts linked by a wide range of conceptual relations different from the traditional hyponymy one. We are currently applying a set of relations (equivalence, location, space, causality, meronymy, etc.) derived from a more general running research on this subject. Also properties can be attributed to a concept.

[1] http://www.geneontology.org/

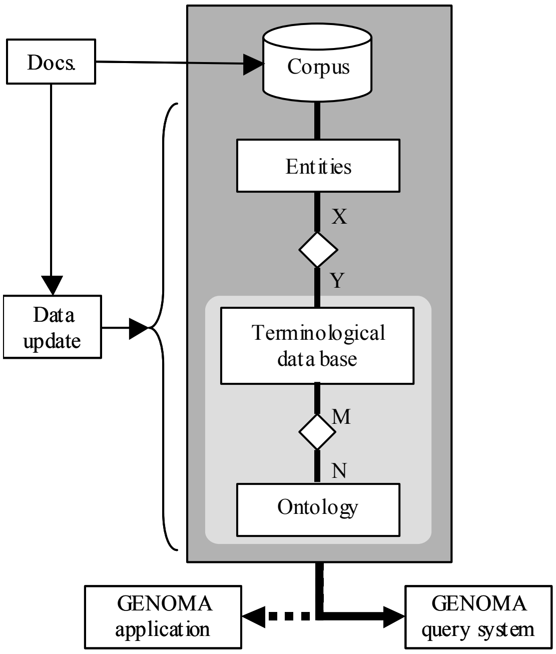
[2] http://www.ncbi.nlm.nih.gov/LocusLink/

[3] http://www.informatics.jax.org/searches/GO_form.shtml

[4] http://bioinfo.weizmann.ac.il/cards/index.html

[5] OntoTerm is a terminological management system built by Antonio Moreno, from the Universidad de Málaga. More information available at: http://www.ontoterm.com.

Figure 1: General structure of the GENOMA-KB



The mutual need between the Term Base and the Ontology

After having reviewed the available resources for terminological management and ontology building, we have decided to use OntoTerm (Feliu; Vivaldi; Cabré, 2002a and b). This tool is based on a conceptual structure previous to the term base creation. In this sense, the ontology building is the previous stage before the construction of the term base. Given this design philosophy, we present firstly the ontology and, secondly, the main characteristics of the term base directly linked to the ontology.

The core ontology was built with the aid of a domain expert who has provided its initial structure for the conceptual structure building. Thus, new concepts have been added to a previous list of base concepts necessary for the system performance. As a matter of fact, the system includes 21 base concepts (ALL, OBJECT, EVENT, PROPERTY, etc.). In addition, the domain expert has proposed a list of about 100 concepts used in the human genome domain that have been integrated to the initial list.

Once the concepts recommended by the domain expert were included in the ontology, the following step consisted on filling in the term base. For each term (related to a given concept) and each language it has been provided with the following specific information:

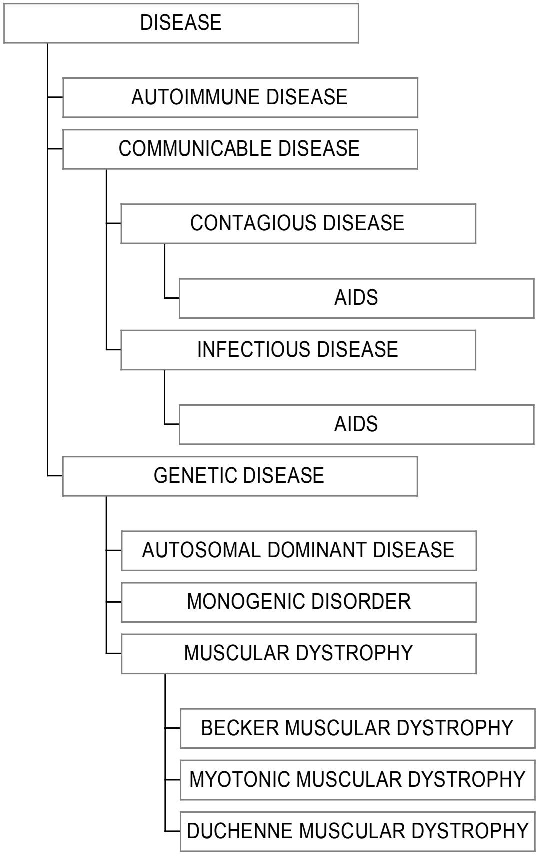
- The term itself
- Part of speech
- Number and gender assignment
- Contexts (each entry term has to have at least one context)
- Context's source
- Lemmatized form (from the morphological point of view)
- Administrative information

All the information above mentioned is mandatory. Apart from it, there is some additional information that is optional: the term definition as well as its source and some usage notes.

The term base methodology consists of selecting *a priori* some tied related concepts forming a category. For example, under the category DISEASE it can be found different types of diseases such as: AUTOIMMUNE DISEASE, COMMUNICABLE DISEASE, GENETIC DISEASE, etc. Similarly, deriving from a given type of disease it can also be found its corresponding subordinate concepts. Thus, for GENETIC DISEASE will appear concepts such as AUTOSOMAL DOMINANT DISEASE, MONOGENIC DISORDER and MUSCULAR DYSTROPHY. Under the term MUSCULAR DYSTROPHY will appear concepts like BECKER MUSCULAR DYSTROPHY, DUCHENNE MUSCULAR DYSTROPHY, MYOTONIC MUSCULAR DYSTROPHY, etc. See Figure 2.

Another important ontology feature is the multiple inheritance. As usual, for multiple inheritance we understand that one concept might have two or more ancestors. This fact can be clearly observed in the case of the concept AIDS, which is at the same time a CONTAGIOUS DISEASE and an INFECTIOUS DISEASE. We think that a robust ontology should account for this kind of phenomena.

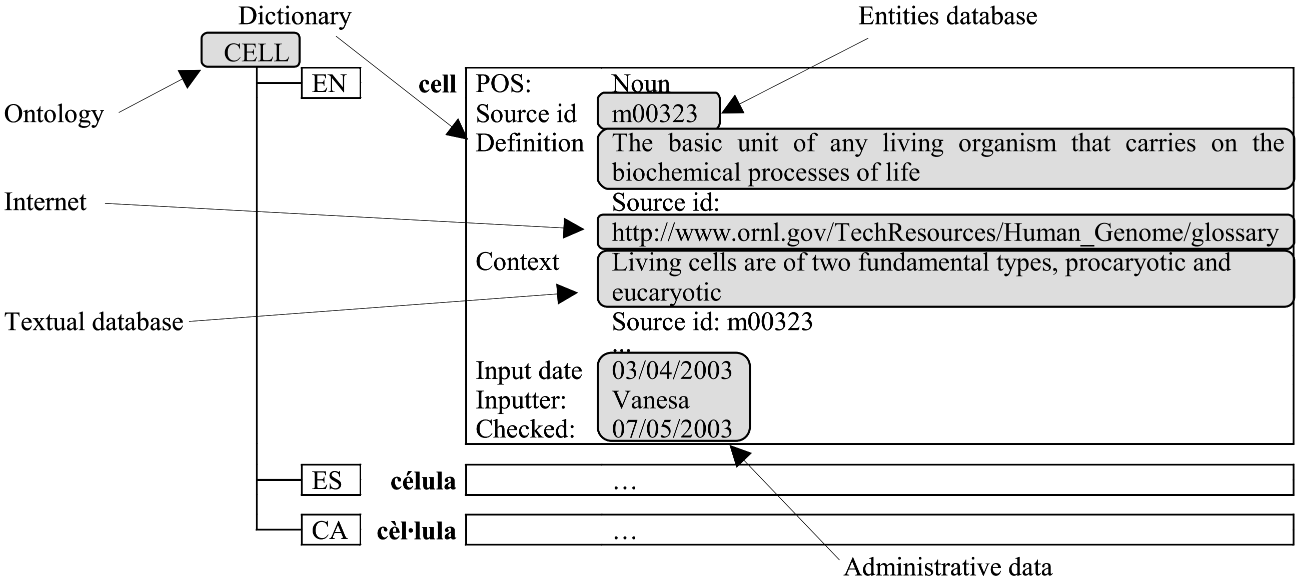
Figure 2. Extract from ontology tree



From the terminologist viewpoint, working on conceptual categories, as for example all concepts related with DISEASE, has two main advantages. On the one hand, it permits to gain efficiency and internal coherence since it becomes a systematic work. On the other hand, it also allows getting cognitive competence in the subject matter.

Now, we describe the criteria applied to retrieve relevant data from the corpus module, which will be the source of information for many of the fields characterizing each term.

Figure 3. Data sources for the term database



As mentioned above, the work is done in a systematic way. Consequently, it is divided into two main steps. Firstly, equivalent terms and their variants are assigned to each ontology concept. Secondly, all the information related to each term entry is added.

In regard to the first step, it is composed of the following phases:

1. Introduction of equivalent terms for Catalan, English and Spanish
2. Selection of variants for each term. The definitions and specially the contexts are the main source for obtaining these variants. The following is an example that illustrates well the way we detect term variants.

   – *"**Duchenne's muscular dystrophy**, the most common and severe type of pseudohypertrophic muscular dystrophy; chronic and progressive, it begins in early".* Called also *Duchenne's d., Duchenne's* or *Duchenne-Griesinger disease, Erb's atrophy* or *Erb's d.,* and *Zimmerlin's atrophy.* Cf. *Becker's muscular d.".*

   – *"Some communities have begun screening for **Duchenne muscular dystrophy (DMD)** by measuring creatine kinase levels in newborns".*

   It is important to remark that not any variant is prioritized. Hence, our work is not prescriptive but descriptive. In fact, our departure point is the text.

Once the first step has been accomplished, it is necessary to enter the remaining linguistic information as well as the administrative information. The procedure is systematic and is reflected in the following stages:

1) *Part of speech* (adjective, noun, verb).
2) *Gender* (specified only for Spanish and Catalan).
3) *Number* (only for lexicalised plurals).
4) *Lemmatized form*. It is used for establishing the link between the term and the units contained in the textual database.
5) *Definition and its source*. It is taken from both analogical and digital specialized dictionaries.
6) *Contexts and their source*. They are taken from the IULA's technical corpus as well as Internet. In the last case, the context's source documents must accomplish some requirements. That is, the information must come from relevant journals, and/or public and private research centers web sites. In addition, these sites must contain all the data dealing with the author, place and year of publication.
7) *Source of the term*. The source comes from the most representative context.
8) *Notes*. They are used generally for remarking preferred uses or irregular forms. For example, *pulmonary alveolus. (Note: usually is documented in plural form. The plural form is "pulmonary alveoli)".*
9) *Inputter*. The name of the person responsible for the entry.
10) *Input date*. The date in which the entry was created.
11) *Check date*. The date in which the entry has been revised.

Figure 3 above shows all the just described information for the English term of the record corresponding to CELL concept.

The Term Base enlargement

As for the second aim is concerned, two different tools, *Mercedes* and *YATE* will be presented. They have been developed in the Institute for Applied Linguistics, and they are used to retrieve terminological units contained in specialised texts, more specifically, in human genome domain texts.

The first tool, *Mercedes*, is a term detector used to retrieve terminological units from texts. Essentially, the system compares the units of a particular domain contained in an internal term database with all the units of a given text. This database has been built from public domain specialised dictionaries and can be easily adapted to any new domain.

YATE is a term candidate extractor based on the combination of different (linguistic and stochastic) strategies. It has also been applied to the same texts in order to obtain a different list of single and multiword terminological units.
Next step of the working process consists of the straight inclusion to the term base of the list of terms containing the coincident units derived from both tools. Following this procedure is mandatory to check that the concept it represents is already included in the ontology. Otherwise the conceptual structure must be updated.
For the remaining list of terms proposed just by *YATE*, it envisages to check their termhood by looking to relevant resources combined with a domain expert consultation. Table 1 shows a contingency table sample of the term candidates proposed by both tools and for each one separately. This information, as already mentioned, will be used for updating both the Term Base and the Ontology.

Table 1. Term Candidates comparison

	Mercedes	non-Mercedes
Yate	*genoma* *gen* *cromosoma*	*mioglobina* *apolipoproteína* *mucina*
non-YATE	*hebra* *hélice* *secuencia*	*microarray* *melanocortina* *retropseudogen*

Conclusions and Future Research

In this paper, we have described the main features of the GENOMA-KB, understood as a multi-module knowledge base integrating the corpus, the bibliographic data, the term base and the directly related ontology. It has been highlighted the methodology followed in order to enlarge the term base and the ontology and the reuse of the results of a term detector and a term extractor to ease the updating task.
In order to reuse the maximum information obtained from these tools it is foreseen, in a running research, to take profit from the textual fragments containing terms (both validated and candidates) linked by a conceptual relation expressed by a verbal form (Feliu, 2004).

Acknowledgements

The authors acknowledge the contribution of Eva Valero in the ontology building.
This research and the development of the query system have been carried on under the two finished public funding projects TEXTERM: *Textos especializados y terminología: selección y recuperación automática de la información* (BFF2000-0841), lead by M. T. Cabré; and RICOTERM: *Sistema de recuperación de información con control terminológico y discursivo* (TIC2000-1191), lead by M. Lorente. And also under the current public funding of TEXTERM2: *Fundamentos, estrategias y herramientas para el procesamiento y extracción automática de información especializada* (BFF2003-2111), lead by M. T. Cabré.

References

Condamines, A; Rebeyrolle, J. (2000). Construction d'une base de connaissances terminologiques à partir de textes: expérimentation et définition d'une méthode. In: Charlet, J. et al. (eds.) Ingénierie des Connaissances, évolutions récentes et nouveaux défis (pp. 225-242). Paris: Eyrolles.

Feliu, J. (2004). Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica. PhD dissertation. Universitat Pompeu Fabra. Barcelona.

Feliu, J.; Vivaldi, J.; Cabré, M. T. (2002a). «Towards an Ontology for a Human Genome Knowledge Base». LREC2002. Third International Conference on Language Resources and Evaluation. Proceedings (pp. 1885-1890). Las Palmas de Gran Canaria, may 2002.

Feliu, J.; Vivaldi, J.; Cabré, M. T. (2002b). Ontologies: a review. Working Paper, 34. Barcelona: Institut Universitari de Lingüística Aplicada.

Martin, L.E. (1990). Knowledge Extraction. In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society (pp. 252--262). Hillsdale, NJ: Lawrence Erlbaum Associates.

Meyer, I.; Bowker, L.; Eck, K. (1992). Cogniterm: An Experiment in Building a Terminological Knowledge Base. Proceedings 5th Euralex International Congress on Lexicography. Tampere, Finland.

Vivaldi, J. (2001). Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD dissertation. Universitat Politècnica de Catalunya. Barcelona.

KB-N: Computerized extraction, representation and dissemination of special terminology.

Magnar Brekke

Department of Professional Communication
Norwegian School of Economics and Business Administration
Helleveien 30, N-5045 BERGEN, NORWAY,
Magnar.Brekke@nhh.no

Abstract

This paper reports early results of a 3-year project aiming to establish a knowledge-bank for economic-administrative domains. Special knowledge is embedded in text produced typically by experts, captured in language independent concepts as language specific terminology which is stratified with respect to domain specificity ranging from general shared terms to unique domain-focal terms. KB-N refines and integrates computational strategies and tools in NLP for corpus design and analysis, automatic and semi-automatic extraction, representation, and retrieval of terminology, dynamic thesaurus creation, dynamic display of authentic collocational and phraseological evidence, etc. In Phase I of the project we are capturing introductory textbook text across 30-odd subdomains. Texts are XML-coded and POS-tagged, and strictly parallel texts aligned for equivalence mining. Term extraction from English text exploits System Quirk's built-in functions while term extraction from Norwegian is being developed from scratch. Pruning/supplementation of candidate lists require man/machine interaction where expert knowledge intersects with terminological principles. The system allows dynamic development of conceptual hierarchies. A range of applications are envisaged for the knowledge bank. The theoretically most interesting use of the KB-N Termbank will be in the context of Norwegian-to-English automatic translation On the didactic side KB-N will be integrated with an established e-learning system.

The concept of a text-based knowledge-bank builds on the underlying assumption that domain-focal special knowledge is embedded in text produced typically by domain experts for documentary, argumentative, didactic or general communicative purposes. It further assumes that the essential knowledge content is embedded in relatively language independent concepts and manifested through relatively language specific terminology (in casu English and Norwegian used in economic-administrative domains), and that such terminology is stratified with respect to domain specificity ranging from general shared terms down to a small set of domain-focal terms.

1. The LSP background

The scholarly study of special languages (LSP, *Fachsprache*, *Langue de specialité*, cf. Sager et al., (1980)) can be traced back to the 20's and 30's, with the Austrian engineer Eugen Wüster's founding of *die allgemeine Terminologielehre* (Wüster, 1932/-1985) as an academic discipline forming a definitive starting-point. The initial development took place largely in the German-speaking areas or former east-bloc countries.

With the post-WWII ascendance of English as the language of science and engineering accompanied by the flocking of students from many countries to universities and professional schools in the UK and the US, a didactic strain of LSP teaching developed which remains a strong Anglo-American tradition – ESP or "English for specific purposes," a subset of EFL/ESL or "English as a Foreign/Second Language". In other regions, especially in parts of Europe, LSP practitioners have developed other research paradigms involving their own native languages and have influenced both research goals and methodology in broad areas of language study and linguistic research

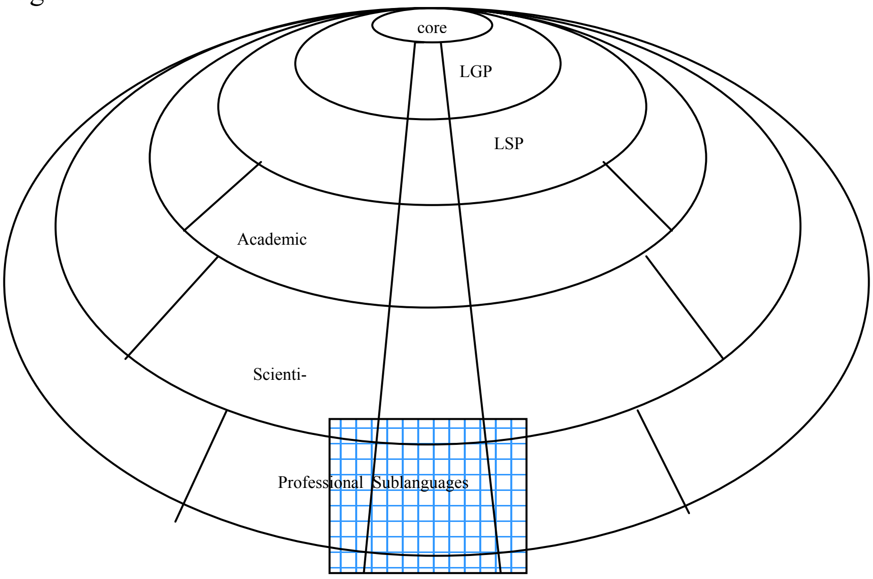
However, despite the overwhelming volume of production and translation of LSP-type text which is actually carried out every year - product documentation and specification, international agreements, treaties and reports, global newsfeed, subtitling and advertising, just to name a few -, to this day LSP has remained a marginal patch in the general field of language and linguistic studies, as evidenced by the relative paucity of dedicated professional journals and conferences.

2. On the nature of sublanguages

It was not incidental that the very first application beyond pure number crunching that was envisaged for the newly invented computer should be the automatic translation of natural languages (Hutchins, 1986), and that the initial strategy involved a massive accumulation of bilingual glossaries of "technical terminology" for automatic look-up – which remains a cornerstone of SYSTRAN, probably the most widespread general MT system to this day (Toma, 1976/1989). The achilles' heel of MT turned out to be the multiply ambiguous general content words of ordinary language (which added to the function words probably make up about 60% of any text). The special terms, however, which convey key concepts of linguistically embedded expert knowledge, in the context of a highly constrained morphology, phraseology and syntax (LSP, in other words, see next section), continue to hold out promises for the processing of natural language through technological applications like MT, e-glossaries, e-learning, e-lexicography etc.

I will now describe the fundamental notions inhabiting this universe of LSP discourse as they are manifested through their linguistic representations. Consider the figure given as diagram 1.

Diagram 1



The figure represents a crude attempt to illustrate the ever increasing "reach" of a language as it is extended from the general sphere shared by nearly all mature native speakers (core + LGP) through increasingly specialized usage strata down to the highly constrained sphere of expert communication. "Core" represents the common core of grammatical elements without which the language simply cannot function, with LGP containing the standard constructions and everyday lexical words which are part of every adult native speaker's linguistic arsenal. LSP is further tentatively stratified as follows: "Academic" suggests a stratum of standard intellectual usages, further narrowed down in the highly goal-focused "Scientific" usage of the various research communities, ending in one of many professional subdomain languages at the bottom of the diagram. Needless to say the borders between any two strata are assumed to be fuzzy, vague and irregular, which diagram 1 cannot convey.

The sector starting at "core" and cutting across all the increasingly specialized strata, including all professional sublanguages, serves to illustrate the simple fact that even the most highly specialized text will necessarily contain the lexis and grammar of ALL the above strata, and dominantly those of core + LGP (as already alluded to). Many of the defining characteristics of a specific LSP text are hence to be sought near the periphery of the stratified model given as diagram 1, where perhaps only a small handful of features and specialized terms will serve to set it apart from those shared by many or all (subdomain) texts.

3. Characteristics of professional sublanguages

The chief burden of specialist communication will thus no doubt be borne by the special lexical units, which I have elsewhere designated "domain-focal terms" (Brekke 2000, 2003). Before delving deeper into the nature of such representations it would be worth taking a few moments to consider some important general characteristics of professional (LSP) communication, covering all strata of diagram 1 outside the core+LGP-portion at the top. Some essential linguistic features can be enumerated (adapted from Brekke, forthcoming):

- the compactness of expression afforded by domain-specific terminology (as manifested through acronyms and abbreviations, composite content words, compound nouns and complex noun phrases)
- occurrence of certain specific constructions, collocations and phraseological conventions
- the utilization of special symbols, formulae and nomenclatures
- over-representation of certain verb types, tenses, "academic hedges" and connectives
- over-representation of impersonal constructions
- under-representation or absence of personal pronouns, emotive adjectives, etc.
- utilization of extra-verbal, semiotic and/or multimedial representations (tables, graphs, diagrams, sound, pictures, animation, video, etc.)

While very few professional languages will be found to exhibit all of these features it would seem equally unlikely to come across one that has none. The compactness of terminology, on the one hand, creates a high degree of precision in content and expression, which is relatively easy to mirror in a target language with a symmetrically developed terminology; otherwise not. On the other hand, the tight specific constraints imposed on collocations and phraseology of a special subdomain language may differ considerably from one national language to the next.

4. On the nature of scientific knowledge

Since the discourse type examined here is of the LSP variety, certain general characteristics follow from the fact that a central function of such languages is knowledge representation. Scientific knowledge on this view is the preliminary result of a largely cumulative process through which it is **established**, that is, discovered, challenged, revised, rejected or passed on. It is then crucially dependent on being **managed**, that is, (linguistically) represented, stored, documented, enhanced, extended, applied etc., in short, fixated in some enduring form. The fate of the ancient Alexandria library testifies to the importance of such careful knowledge management, as does the ominous disintegration of 20-year old CD-ROMs. But a third dimension is equally important: the knowledge so conserved will be of little use unless it is also **conveyed**, that is, taught, disseminated, published, communicated in some form or channel. The history of learning abounds in examples of discoveries or inventions that went unpublished and had to be repeated for this reason.

A professional **text**, under this view, is seen as embodying a subset of the universe of knowledge established along the lines sketched above. The gateways to this universe of knowledge are made up of **concepts** (ideas, notions, "units of thought") which have defined relations (generic, partitive and so on) to other concepts, and often a defined hierarchic position based on some ontology.

Concepts are in principle language-independent, i.e. elements of a cognitive interlingua, although the human mind depends on some *signifiant* providing mnemonic anchorage, otherwise the saussurean sign will fall apart. At the meta-level we resort to recognizable words in identifying a given concept, such as SOLSTICE or VALUE or GRAVITY (notice that for the capitalized items to "make sense" the reader is dependent on knowing what the corresponding English words "mean" – a number like 2386 would not work). In principle the conceptual gate remains closed until we find the appropriate **term** which can unlock the gate and access the meaning content which is being singled out – or conversely, the scientist having just made a discovery or invention, completed a mathematical proof or drawn a crucial inference will be groping for a way to break his or her abstract idea out of its cognitive prison and represent it through a concrete linguistic element. From this brief sketch of the theoretical and methodological underpinnings we move now to the practical applications.

5. Implementations

5.1. Modules and functions

KB-N represents the culmination of efforts to refine and integrate computational strategies and NLP tools for

- corpus design and analysis
- automatic and semi-automatic extraction, representation, and retrieval of terminology
- dynamic thesaurus creation
- dynamic display of authentic collocational and phraseological evidence

The paper will attempt to demonstrate (if possible on-line, if not via screen-shots) the working version of the integrated KB-N software suite for handling corpus based concordancing, term extraction and selection, thesaurus building on-the-fly etc., in the context of a discussion of the above general theoretical and methodological issues.

An elaborate semi-hierarchic classification of essential subdomains of economics and business administration forms the basis of storage and retrieval of terms as well as texts. Since it is concept-based the term-bank will easily accommodate the addition of more languages (German, French and Spanish are being contemplated).

5.2. Textual basis

In Phase I of the project we are capturing introductory textbook text across 30-odd subdomains; research articles and popularization will be added later. Texts are scanned or downloaded from relevant sources, routinely XML-coded and POS-tagged using Stuttgart's Corpus Workbench. Purely non-language material is deleted. Textbook indexes and existing glossaries are exploited for definitional purposes, complemented by established references. Where strictly parallel texts are available the two language versions are aligned for equivalence mining. For the text-bank a variety of restrictions apply; depending on the copyright status of the material, the end user will be allowed to see a concordance line, a paragraph, or a screenful.

5.3. Term extraction

Term extraction from English text exploits System Quirk's "Weirdness"-function (comparing the occurrence ratio of a given item in a text with that of a general reference corpus) as well as its "Ferret" module (which identifies strings of lexical words uninterrupted by function words); taken together these produce reasonably good term candidate lists once the stop-lists have been refined.

Term extraction from Norwegian text is being developed through an associated project. We have access to Hofland's massive corpus of Norwegian newspaper text (http://helmer.aksis.uib.no/aviskorpus/) as a standard of comparison for "weirdness"-studies, while concatenation of lexical roots (which Norwegian shares with other "pure" Germanic languages) requires other search tactics than Quirk's "ferreting".

5.4. Concept systematization

Getting from automatically generated term candidates to solid terms is a critical stage where domain expert knowledge intersects with terminogical principles, and man/machine interaction must lean heavily on the former, especially in identifying "missing" concepts. The same holds for the establishing of conceptual hierarchies: As explained above concepts are viewed as gateways to the domain knowledge, and the way they are organized and structured requires deep understanding of the domain. The KB-N system provides essential support in allowing dynamic conceptual hierarchies to be developed, changed and revised alongside the terminological analysis.

5.5. Applications

A range of applications are envisaged for the knowledge bank. It is designed as a web-enabled resource available for systematic terminology registration and look-up, textbook authoring, e-learning as well as machine translation. The theoretically most interesting use of the KB-N Termbank will be in the context of Norwegian-to-English automatic translation of the relevant types of domain text. On the didactic side KB-N will be closely integrated with an established e-learning system to provide interactive study support for students of economic-administrative subjects.

References

Ahmad, K. (1996). "A terminology dynamic and the growth of knowledge: a case study in nuclear physics and in the philosophy of science." In *Proceedings of the 4th International Congress on Terminology and Knowledge Engineering, Vienna*: 1-11. Frankfurt: INDEKS-Verlag.

Brekke, M. (2000) "Lexical Identification of Domain Focal Text and Terminology", paper read at COMLEX 2000, Patras, Greece.

Brekke, M. (2002) "TERMINEC. A Clearinghouse for Economics Text and Terminology", in Peters, P., Collins, P & Smith, A, (eds) *New Frontiers of Corpus Research. Proceedings of ICAME 2000.* Amsterdam: Rodopi.

Brekke, M.(forthcoming) "Linguistic aspects of the translation of scientific and technical text", in Kittel, H. et al., *Translation: An International Handbook of Translation Studies.* Berlin: de Gruyter,.

Hofland, K. Norwegian Newspaper Corpus:http://helmer.-aksis.uib.no/aviskorpus/.

Hutchins, J. (1986) *Machine translation: past, present, future.* Chichester: Ellis Horwood.

Lehrberger, J. and L. Borbeau (1988). Automatic translation and the concept of sublanguage. In Kittredge, R.I. and J. Lehrberger (eds.). *Sublanguage: studies of language in restricted domains*: 81-106. Berlin: de Gruyter

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* 252--262. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sager, J.C., D. Dungworth, P.F. McDonald (1980). *English special languages.* Wiesbaden: Brandstetter.

SystemQuirk: http://www.mcs.surrey.ac.uk/SystemQ

Toma, P (1976/1989) An operational machine translation system. In Chesterman, A. (ed) *Readings in translation theory*: 162-172. Helsinki: Finn Lectura.

Wüster, E. (1932/1985) *Einführung in die allgemeine Terminologielehre und terminologische Lexicographie.* 2nd ed. Copenhagen: Fagsprogligt Center, CBS.

Acknowledgements:
The research reported here is supported by the Norwegian Research Council through its Language Technology Program (KUNSTI) as well as by the Norwegian School of Economics and Business Administration (NHH). KB-N is based in NHH's Center for LSP Research in cooperation with UiB's Aksis-centre and Fagbokforlaget.

The Cause-Effect Relation in a French-Language Biopharmaceuticals Corpus: Some Lexical Knowledge Patterns

Elizabeth Marshman
Observatoire de linguistique Sens-Texte (OLST)
Département de linguistique et de traduction, Université de Montréal
C. P. 6128, succ. Centre-ville
Montréal, Québec
CANADA H3C 3J7
elizabeth.marshman@umontreal.ca

Abstract

This project used two corpora of texts on biopharmaceuticals, one in English and one in French, to study lexical knowledge patterns i.e., lexical units or combinations of lexical units often used in association with a conceptual relation, in this case the cause-effect relation. The purpose of this research was to identify a selection of knowledge patterns that would be useful for identifying knowledge-rich contexts (KRCs) in text corpora. This paper discusses some of the French patterns found and compares the results and the difficulties encountered to those observed in the English corpus. In particular, the issues of pattern polysemy, variation and repetition, hedging, and the presence and usefulness of cognate pairs in the patterns in English and French will be discussed.

1 Introduction and Goals

This paper describes the French-language portion of a Master's-level research project carried out at the University of Ottawa (Marshman, 2002a). It is the counterpart of an article on the results of the English-language research (Marshman, 2002b).

The general purpose of this research project was to produce a list of lexical knowledge patterns that can enable users to automatically and precisely locate and extract knowledge-rich contexts (KRCs) from text corpora. This article describes some of the more precise patterns found for French and compares these patterns and some of the issues observed in their identification and application to those observed in the English corpus.

2. Some Basic Concepts

2.1 The Cause-Effect Relation

The cause-effect relation is one of the most fundamental in human perception, and has a long history of study in various fields. It is especially important in the sciences and in medicine in particular; the basic purpose of medicine is after all to identify causes and mechanisms of health and disease and to treat and prevent illness. However, the cause-effect relation has been studied by a relatively small number of researchers in a terminological context, among them Nuopponen (1994), Cabré et al. (1996), Garcia (1996, 1997) and Barrière (2001, 2002). Cognitive scientist Leonard Talmy (1988) has also produced a very interesting description of the relation. Barrière (2002) took Talmy's description as the basis of her classification, which was used in turn for this project.

Talmy defines the relation through force dynamics, in terms of the interaction between two opposing forces, an agonist and an antagonist. The sub-classification of the relation is determined by the result of the interaction between the two forces. Barrière, following Talmy, describes two major subdivisions of cause-effect relationships, which each in turn have four divisions. The *existence dependency* describes cases in which the interaction of forces defines the existence of the effect, and includes sub-relations of creation, destruction, maintenance, and prevention. In the *influence dependency* the interaction affects a characteristic or quality of the effect, and includes sub-relations of modification (in which there is an undefined effect), two sub-types of modification, increase and decrease, and preservation.

2.2 Lexical Knowledge Patterns

Knowledge patterns can be defined as textual elements that are often associated with a certain conceptual relation (Meyer, 2001; Marshman, Morgan and Meyer, 2002). Lexical knowledge patterns, the focus of this research, consist of lexical units or groups of units. Some examples of lexical patterns are *(is a) type of* for the generic-specific relation and *(is a) part of* for part-whole.

The approach here focused on automatic extraction of contexts followed by manual analysis of the results. With rare exceptions, the patterns identified are simply character strings; no preliminary grammatical or semantic analysis of the texts was performed.

Knowledge patterns can be used in several ways. Among them is the eventual goal of this research: incorporation into knowledge extraction tools. These tools can be used to extract from electronic corpora contexts describing concepts related to each other in specific ways, i.e., in this case, cause and effect. For example, using such a tool a user (e.g., a terminologist) may search for a term appearing in proximity to one or more cause-effect patterns. The tool will then extract contexts likely to provide information about the cause-effect relations involving the concept designated by the term.

2.3 Evaluation of Knowledge Patterns

The study of knowledge patterns requires evaluation of the likely usefulness of the patterns in application. This project used precision as the primary measure for evaluation, followed by frequency.

2.4 Corpus and Domain

Our corpus, approximately 224,000 words in length, was composed of French-language texts from the subject field of biopharmaceuticals. These included general texts (e.g., newspaper and magazine articles), specialized texts (e.g., abstracts of research papers, drug company information), and some technical texts (e.g., full-text research papers, prescribing information for physicians).

3 Methodology

This project was carried out in six main steps: 1) the building of the corpora from texts available on the Internet and from electronic databases; 2) the generation of concordances for domain terms; 3) the manual analysis of these concordances to identify potential lexical knowledge patterns; 4) the generation of concordances using these candidate patterns; 5) the manual analysis of the concordances to classify the contexts identified by their usefulness for knowledge extraction; and 6) the calculation of the patterns' precision in the corpus.

The contexts were classified into three groups, depending on whether: a) the relation was clearly present with both concepts overtly indicated (hits); b) the relation was present but less clearly, or was present but with one of the concepts not clearly indicated ('maybes'); or c) the relation was not present or the sentence was unusable for knowledge extraction (noise). Precision for each of the patterns was calculated using the definite hits alone, and also for the superset of definite hits and 'maybes'.

4 Knowledge Patterns Observed

This section presents a selection of the more precise knowledge patterns observed in the corpus, accompanied by an indication of their precision in the corpus, and with examples of each of these patterns in use.[1]

4.1 Existence Dependency

4.1.1 Creation

provoc/provoq** (100/100) [*provoke*]

De plus, ce traitement **provoque** des effets secondaires souvent très gênants.

*parce qu** (100/100) [*because*]

Les diabétiques ne peuvent assimiler correctement le sucre (glucides) **parce que** leur pancréas ne sécrète pas correctement une hormone qui active l'utilisation du glucose dans l'organisme.

*entraîn** (98/100) [*lead to*]

L'administration d'une suspension orale de charbon activé en poudre ou de colestyramine **entraîne** une augmentation rapide de la clairance plasmatique de l'A771726 (voir section surdosage); il en résulte une réduction de la demi-vie d'élimination à 24 heures.

*stimul** (86/100) [*stimulate*]

Insuline de séquence humaine, moins immunogène que les insulines bovines ou porcines. Se lie à la sous-unité cellulaire de son récepteur et **stimule** l'activité tyrosine kinase de la sous unité bêta par autophosphorylation.

4.1.2 Destruction

nocif/nocive** (100/100) [*harmful*]

... plusieurs effets **nocifs** qui sont attribués aux statines pourraient être reliés à leur activité dans les tissus extrahépatiques.

interfér/interfèr** not *interféron* (100/100) [*interfere*]

... l'inhibition de l'OSC n'**interfère** pas avec leurs synthèses respectives...

supprim/*suppress** (83/100) [*suppress*]

… la synthèse de CH est **supprimée** par une inhibition de l'enzyme hydroxyméthylglutarylcoenzyme A (HMGCoA) réductase, qui est l'enzyme déterminant la vitesse de réaction…

tue/tué** (71/100) [*kill*]

... les antibiotiques diminuent l'effet **tueur** du virus sur des cellules infectées.

4.1.3 Maintenance

permet/permi** (97/99) [*allow, permit*]

Le séquençage du gène codant pour la TK ou pour l'ADN polymérase **permet** l'identification du support génétique de la résistance.

indispensable (pour/à/au)* (70/100) [*indispensable* (*for*)]

Ce sont des antiprotéases, substances qui bloquent une enzyme **indispensable à** la fabrication du virus.

*nécess** (56/92) [*necessary*]

La mesure in vitro de la sensibilité des HSV aux antiviraux est réalisable en routine par des techniques simples comme le test colorimétrique au rouge neutre, mais elle **nécessite** l'isolement préalable du virus sur cellules.

*essentiel** not *essentiellement* (53/86) [*essential*]

L'hormone de croissance est **essentielle** pour la protéine de synthèse, l'action hormonale (spécialement les hormones pour la thyroïde et le sexe), une récupération après l'exercice...

4.1.4 Prevention

empéch/empêch** (94/98) [*prevent*]

... une multithérapie **empêcherait** le virus d'infecter d'autres cellules...

bloqu/blocage** (90/100) [*block*]

De façon similaire, l'acivicine **bloque** la croissance des cellules du lignage B.

*enray** (90/100) [*eliminate*]

On doit maintenant vérifier si cette protéine ainsi «déguisée» permet d'**enrayer** la capacité du virus à se reproduire dans de vraies cellules infectées, prélevées chez des individus séropositifs.

4.2 Influence Dependency

4.2.1 Modification

*influenc** (100/100) [*influence*]

Les grandes variations dans l'activité de ces enzymes peuvent **influencer** la réponse individuelle à la lovastatine.

*neutralis** (95/100) [*neutralize*]

En fait, notre idée est de se servir de la protéine Vpr comme cheval de Troie pour introduire un agent antiviral dans le virus pour le **neutraliser**.

[1] In the pattern forms, truncation is indicated by an asterisk (*), an alternative by a slash (/), exclusion by *not*, and optional elements by parentheses. The precision (%) observed in the corpus is indicated in parentheses after the pattern form, with the precision of hits exclusively first, and the precision including the 'maybe' category second. A possible English equivalent of the pattern is provided in square brackets after the precision. Within the contexts, the cause is indicated by single underlining, the effect by double underlining, and the pattern form in bold.

(s')impliq dans* (94/100) [(*be*) *involved in*]
... il y a un lien entre la synthèse du cholestérol et l'activité de la 7ahydroxylase (l'enzyme **impliqué** dans la formation d'acides biliaires) dans le foie.

*(jouer) rôle** (+ adjective) (78/99) [(*play a*) *role*]
Acide ascorbique (COH) : **joue un rôle** très **important** en assurant la régénération de l'atocophérol en se transformant en un radical très peu réactif (CO·)...

4.2.2 Increase

*contribu** (95/100) [*contribute*]
La réabsorption accrue de sels biliaires tend à contrecarrer l'effet bénéfique de la thérapie aux statines et **contribue** à la prolongation du temps nécessaire à l'obtention de l'état d'équilibre.

*favoris** (92/96) [*favour*]
De fait, une augmentation du contenu en cholestérol des plaquettes **favorise** l'activation des plaquettes chez les patients hyperholestémiques.

*augment** (64/86) [*increase*]
De plus, la température élevée **augmente** l'efficacité de plusieurs globules blancs et de certaines substances antivirales.

*amélior** (60/100) [*improve*]
Le traitement substitutif par hormone de croissance est le seul recours pour **améliorer** la taille des enfants ayant un déficit somatotrope avec un bon rapport efficacité/tolérance.

4.2.3 Decrease

*ralenti** (80/100) [*slow*]
L'hormone de croissance augmente les masses protéiques au niveau musculaire, des organes et du sang tout en **ralentissant** la destruction des protéines et en activant leur production.

*diminu** (76/94) [*diminish*]
… le retrait de LDL est **diminué** par une baisse de production de récepteurs LDL…

abaiss/baiss** (61/99) [*lower, decrease*]
Il y a cinq ans, la mortalité atteignait 95 %. La découverte d'un agent antiviral l'a **abaissée** à 30 %.

*inhib** (60/100) [*inhibit*]
L'amantadine et la rimantadine … **inhibent** également de façon non spécifique la multiplication des virus grippaux de type B et C ainsi que d'autres virus tels que le virus respiratoire syncytial (VRS), les virus parainfluenza 1, 2 et 3 ou encore les virus de la dengue…

4.2.4 Preservation

mainten/maintien** not *maintenant* (60/100) [*maintain*]
L'effet combiné de ces deux actions est de **maintenir** de faibles taux sanguins de LDL.

5 Comparison to Previously Examined English Patterns

While the scope of this project limited the comparative analysis that was possible, striking similarities were found in both overall precision and distribution of the patterns (in terms of total frequencies and of the grammatical categories to which they belonged) in English and French (cf. Marshman, 2002b). Perhaps the most interesting is the similar lack of precise patterns for the sub-relation of preservation. Only one pattern in French and none in English produced more than 50% precision in the corpus. The source of this lack remains to be determined, but could be related to the importance of the sub-relation in the field, to the classification of the occurrences, or to the observation process itself.

It is worth noting that other projects focusing on markers of the cause-effect relation have focused on verbs exclusively (Garcia 1996, 1997) or primarily (Barrière 2001, 2002). While in both English and French a large proportion of the patterns observed in this project were verbs, other parts of speech were found. In an analysis of the patterns observed, verb forms (including past and present participles) accounted for 54% of pattern occurrences in the English corpus, noun forms (including nominalizations of verbs) for 32%, and adjective forms for 14%. In the French corpus, verb forms accounted for 49% of the pattern occurrences, noun forms for 44%, and adjective forms for 7%. A similarly small variation was found in a comparison of the proportions of pattern forms in the two languages. The higher pattern occurrences in verb form in English and in noun form in French seems to correlate with observations of trends in general language.

Some of the issues and difficulties noted in the process of identifying and applying the knowledge patterns to extract contexts did seem to be particularly closely linked to language (as observed in other projects, e.g., Marshman, Morgan and Meyer 2002). Overall, there was higher pattern variability in French, linked to factors such as number and gender variation of nouns, pronouns and adjectives, more highly inflected conjugated verbs, and diacritics. On average, the English patterns identified in the project occurred in 3.8 different forms, the French patterns in 7.4.

The impact of this variation is relatively low in cases where truncation can be used to find all variants of a form. However, variation multiplies the work involved in creating the list of pattern forms in cases where truncation is not practical, as when there is no stable root (e.g., *générer/génère*) or when a list of individual forms must be used in order to exclude excessive noise from similar word forms (e.g., as was the case in English with *prod/prods/prodding/prodded*, in order to differentiate this pattern from forms of *produce* and especially to exclude *product*, which generated too much noise). In the patterns identified in this project, 19 (18%) of the more precise English patterns and 26 (28%) of the French patterns required that two or more separate forms be specified. Further work to improve pattern precision will lead to even more multi-form patterns.

Increased pattern variation in French could also be linked to factors that are widely recognized in general language, such as lower tolerance for repetition (e.g., the average frequency of the English patterns was 25, while average frequency of French patterns was 20) and more frequent modification of sentence structure for emphasis.

One of the major difficulties in the use of the patterns was their ambiguity. Categorial ambiguity was common in the patterns, with 26 pattern forms in English and 10 pattern forms in French showing noun/verb ambiguity. This suggests that part-of-speech-tagged corpora could be very useful for working with patterns, especially in cases in which a form is more productive and/or precise when associated with a particular part of speech than with another (e.g., *produce* in English is productive as a verb, but not as a noun). However, by far the more complicated problem was polysemy. Many potential patterns, including several prepositions and conjunctions (e.g., *après*, *avec*) are often used in contexts indicating a causal relationship, but are so polysemous they generate far too much noise to be used in this kind of approach. While these patterns may indicate a cause-effect relation or another kind of relation altogether, others may indicate different kinds of cause-effect relations, making it difficult to automatically identify a specific sub-relation using the pattern. This is the case, for example, with the patterns in the category of modification, which may often be used to indicate the more precise relations of increase or decrease. These problems of polysemy are widespread in both French and English, and will at the very least require lengthy analysis of the patterns before the noise they produce can be reduced to acceptable levels and contexts classified automatically by the relation indicated.

Hedging and modals were very frequently used in both corpora. However, the use of the conditional tense for hedging was unique to French. As Barrière (2002) noted, some method for addressing the issue of certainty is necessary for analysis of causal relationships in texts. This might offer a straightforward way to automatically identify (and/or exclude if so desired) some contexts including hedging, as in many cases conditional forms of verbal patterns themselves could be easily targeted.

Finally, the presence of many cognates (English and French patterns which shared the same roots) was noted among the patterns, which suggested at first that cognates might be an interesting starting point for developing parallel lists of lexical knowledge patterns in other languages. However, the cognates observed varied widely in both their frequency and their precision for identifying the relation. Only one-third of the cognates observed showed less than 10% variation in their precision in the two corpora; only one-quarter showed frequencies that varied by less than ten occurrences, and only one-sixth showed both similar precision and frequency. Among this latter group were *incit* / incit** (*à*) [e.g., *incite; inciter*], *confer* / conférer** [e.g., *confer; conférer*], *consequence* / conséquence** [e.g., *consequence; conséquence*], *maximiz* / maximis** [e.g., *maximize; maximiser*], *accelerat* / accél** [*accelerate; accélerer*], *inactivat* / inactiv** [e.g., *inactivate; inactiver*], and *influenc** [e.g., *influence; influencer*]. On a positive note, however, the sub-relation indicated by the cognates was almost exclusively the same in the two languages. From this perspective, testing cognates in other languages may be worthwhile, but the usefulness of the patterns must be tested in each language.

6 Conclusions and Suggestions for Future Research

As in the case of the English patterns, the results of this research were promising. Several very precise patterns were identified, and could be further tested, refined and used as part of computerized knowledge extraction tools. This testing and refinement will be very important in overcoming some of the challenges of this approach (such as high levels of categorial and semantic ambiguity of the patterns identified), and in identifying the real value of the patterns for information extraction. Notably the recall of the patterns will need to be evaluated, and some of the issues of pattern ambiguity and noise addressed.

One approach for dealing with the issues of ambiguity and polysemy could be a more refined, linguistic analysis of the corpora and the patterns, achieved for example by using a part-of-speech tagged corpus and an analysis of the collocational properties of the patterns. However, it will be important nevertheless to remain on a level of analysis that is consistent with automatic applications.

7 Acknowledgements

The author wishes to acknowledge the generous support of the Social Sciences and Humanities Research Council of Canada. She would also like to thank Ingrid Meyer, Lynne Bowker, and Caroline Barrière for their assistance in the course of this research.

8 References

Barrière, C. (2001). Investigating the Causal Relation in Informative Texts. Terminology 7(2), 135--154.

Barrière, C. (2002). Hierarchical Refinement and Representation of the Causal Relation. Terminology 8(1), 91--111.

Cabré, M.T., Morel J. & Tebé C. (1996). Las relaciones conceptuales de tipo causal: un caso práctico. In Actas del V Simposio Iberamericano de terminologia RITerm. Mexico City. [http://www.unilat.org/dtil/MEXICO/cabremt.html] Consulted 19 September 2003.

Garcia, D. (1996). COATIS, un outil d'aide à l'acquisition des connaissances causales exprimées dans les textes. Actes du Colloque Linguistique et Informatique de Montréal, CLIM'96 (pp. 97--103).

Garcia, D. (1997). Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'Action dans les textes. In TIA-97 : Actes des deuxièmes rencontres terminologie et intelligence artificielle (pp. 7--26). Toulouse : Toulouse-le Mirail.

Marshman, E. (2002a). The Cause Relation in Biopharmaceutical Corpora: English and French Patterns for Knowledge Extraction. Master's Thesis, School of Translation and Interpretation, University of Ottawa.

Marshman, E. (2002b). The Cause-Effect Relation in a Biopharmaceutical Corpus: English Knowledge Patterns. In Proceedings of Terminology and Knowledge Engineering 2002 (TKE'02) (pp. 89--94). Nancy, France.

Marshman, E., Morgan T. & Meyer I. (2002). French Patterns for Expressing Concept Relations. Terminology 8(1), 1--29.

Meyer, I. (2001). Extracting Knowledge-rich Contexts for Terminography: A Conceptual and Methodological Framework. In D. Bourigault, C. Jacquemin, and M.-C. L'Homme (Eds.), Recent Advances in Computational Terminology (pp. 279--302). Amsterdam/Philadelphia: John Benjamins.

Nuopponen, A. (1994). Causal Relations in Terminological Knowledge Representation. Terminology Science and Research 5(1), 36--44.

Talmy, L. (1988). Force Dynamics in Language and Cognition. Cognitive Science 12, 49--100.