First International Workshop on Terminology and Lexical Semantics (TLS'09)

Proceedings

edited by

Amparo Alcina

and

Marie-Claude L'Homme

The TLS Workshop is held in conjunction with

Fourth International Conference on Meaning-Text Theory (MTT'09)

Université de Montréal
Montreal (Quebec) CANADA
June 19, 2009

Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal

ISBN 978-2-9811149-1-4

Organizers

Amparo Alcina (TecnoLeTTra Team, Universitat Jaume I, Castellón, Spain)

Marie-Claude L'Homme (OLST, Université de Montréal, Canada)

Program committee

Ignacio Bosque (Universidad Complutense de Madrid, Spain)

Béatrice Daille (LINA-CNRS, Université de Nantes, France)

Patrick Drouin (OLST, Université de Montréal, Canada)

Pamela Faber (Universidad de Granada, Spain)

Kyo Kageura (University of Tokyo, Japan)

Patrick Leroyer (University of Aarhus, Denmark)

François Maniez (CRTT, Université Lumière Lyon-2, France)

Elizabeth Marshman (University of Ottawa, Canada)

Alain Polguère (OLST, Université de Montréal, Canada)

Margaret Rogers (University of Surrey, UK)

Thomas Schmidt (University of Hamburg, Germany)

Carlos Subirats (Universitat Autònoma de Barcelona, Spain)

Carles Tebé (Pompeu Fabra University, Spain)

Leo Wanner (Pompeu Fabra University, Spain)

Pierre Zweigenbaum (LIMSI-CNRS & ERTIM-INALCO, France)

Aknowlegments

Organized with support from Fonds de recherche sur la société et la culture (FQRSC) of Quebec and Département de linguistique et de traduction of Université de Montréal.

Contents

First International Workshop on Terminology and Lexical Semantics (TLS'09)

Foreword

Amparo Alcina

Marie-Claude L'Homme

Recent work on terminology processing (corpus-based or knowledge-based terminology, computational terminology, terminography or specialized lexicography, etc.) has shown the importance of taking into account the linguistic aspects of terms as well as their conceptual aspects. Linguistic aspects are often those that must be dealt with when processing or analyzing corpora. Conceptual aspects, on the other hand, are those that terminologists must take into account when building knowledge representations in specific subject fields. The weight given to each aspect depends on the theoretical model on which a terminology project is based, its objectives or the application for which it is developed.

Although conceptual and linguistic aspects of terms are recognized as being necessary, less work has been carried out on their linguistic properties and/or their formal representation. Which linguistic properties should be represented when faced with a specific application? Which theoretical model is better adapted to terminological needs? How can linguistic properties of terms - especially lexico-semantic aspects - be implemented in terminology applications? All these questions must be addressed and resolved when processing specialized texts and representing knowledge.

Our aim - when we issued the call for papers for this workshop - was to bring together researchers interested in linguistic approaches - especially lexico-semantic aspects - to the description of terms, and provide an opportunity to discuss issues of a fundamental nature (e.g., the modelling of meaning or the definition of sets of lexical relationships) as well as applied work dealing with the acquisition of lexical aspects of terms and/or their representation in formal structures. We believed that organizing the workshop in conjunction with the Meaning-Text Theory (MTT) Conference would provide us with an opportunity to attract researchers interested not only in conceptual aspects of terminology but also in linguistic aspects. We believe we have succeeded since the workshop provides a forum for addressing issues such as processing specialized corpora using theoretical principles based on lexico-semantic as well as concept-based approaches, building monolingual or multilingual dictionaries or databases using different perspectives, building concrete applications for information science or tools designed to assist terminologists in their every day work.

In addition to the eight articles selected for the workshop, we asked Leo Wanner to present an invited talk on how a specific lexico-semantic framework, namely Explanatory Combinatorial Lexicology (ECL), can serve as a sound theoretical basis when dealing with specialized language, especially in the context on natural language processing (NLP). Among other things, L. Wanner shows how two fundamental properties of lexical units - argument structure and lexical relationships - can be modelled formally in ECL. He also discusses some similarities and differences that need to be taken into consideration when processing lexical units from the point of view of general language and from the perspective of specialized language.

The articles collected in this volume present innovative work that offers different perspectives on terminology and lexical semantics. Some work examine how lexical semantics frameworks, such as Semantic Frames (M.-C. Wermuth, F. Hadouche, M.C. L'Homme, G. Lapalme and A. Le Serrec) or, as was said above, Explanatory Combinatorial

Lexicology (L. Wanner) can be used in terminology applications or when processing specialized corpora. Other articles focus on the combination of concept-based and lexico-semantic approaches to terminological descriptions or modelling of specialized meaning (H. Tissari and T. Chezek; F. Maniez). Quite a few articles present work related to natural language processing (NLP) applications and show how terminology can be dealt with in this context. This work addresses issues related to indexing (L. Da Sylva), bilingual extraction (E. Marshman and P. Bolderen), machine translation (L. Wanner), dictionaries or terminological databases (E. Valero, V. Pastor and A. Alcina), automatic annotation (F. Hadouche, M.C. L'Homme, G. Lapalme and A. Le Serrec), computer-aided translation or terminology (E. Marshman and P. Bolderen; E. Macklovitch, G. Lapalme and N. Chan).

We hope this workshop will be the starting point of fruitful discussions on terminology and lexical semantics and how both approaches can be combined and lead to applications that can handle conceptual properties as well as linguistic properties of terms. We would like to take this opportunity to thank the organizers of the main conference for allowing us to devote an entire day to terminology and lexical semantics. We would also like to thank the members of the program committee for the time they spent selecting the articles, the Fonds de recherche sur la société et la culture (FQRSC) and the Département de linguistique et de traduction of the Université de Montréal for their financial support and, finally, Emmanuel Chièze and Benoît Robichaud for their help in various aspects of the organization of the TLS workshop.

Term-spotting with TransCheck: A Capital Idea

Elliott Macklovitch, Guy Lapalme
Laboratoire RALI
Département d'informatique et de recherche opérationnelle
Université de Montréal
macklovi@iro.umontreal.ca
lapalme@iro.umontreal.ca

Nelida Chan
Government Translation Service
Ontario Ministry of Government and Consumer Services
77 Grenville Street
Toronto, ON M5S 1B3
Nelida.Chan@ontario.ca

Abstract

TransCheck, the RALI's automatic translation checker, has recently undergone a field trial at the Government Translation Service of Ontario, where the system was used not only to detect inconsistent terminology, but also to find new source language terms in texts sent to outside translation suppliers. We describe a specialized term-spotting module developed in the course of that trial to assist the terminologists identify new official names to be added to ONTERM, the Ontario government's online terminology database.

1 The TransCheck system

The RALI Laboratory has been developing the TransCheck system (henceforth abbreviated as TC) for some years now.[2] As its name suggests, TransCheck was originally conceived as a translation checker, i.e. as a system that would be used by a translator or a reviser to detect possible errors in a draft translation, somewhat like a spell checker. Unlike a spell checker, however, which detects errors of form in a single monolingual text, TransCheck is designed to detect errors of correspondence that occur between two texts – a source text in one language and its translation in another. So, for example, the French word 'librairie' would not be flagged as an error by a monolingual spell checker, since it is a correct form of the French language; however, that same form could be flagged as an error of correspondence by TC if it appeared in a target text as the translation of the English word 'library'.

Roughly speaking, TC works as follows. The user begins by specifying a source and a target text file, which the system first tokenizes (i.e. segments into words and sentences) and then automatically aligns. The latter is a crucial step in which TC determines which target sentence(s) correspond to each source sentence. To these aligned regions, TC then applies its various error detection modules. The system currently detects the following types of errors: inconsistent terminology, or terms that diverge from those required by the user (which we call the 'positive' terminology); source language interference, including false cognates like 'library/librairie' (called 'negative' terminology); and paralinguistic expressions, like numbers, dates and monetary expressions. Of course, a draft translation may well contain many other types of errors, but to automatically detect these will often require a deep understanding of the two texts that are in a translation relation. For the time being, TC limits itself to the aforementioned set of errors, all of which can in principle be detected by relatively simple, purely formal means. TC flags a potential error when, in an aligned region, it either detects a certain source item (e.g. a numerical expression) without finding its obligatory target correspondent, or when a source item is detected along with its prohibited target correspondent (e.g. a false cognate). After the complete bi-text has been processed in this way, the

[2] For a detailed description of an earlier version of TransCheck, see (Macklovitch, 1994).

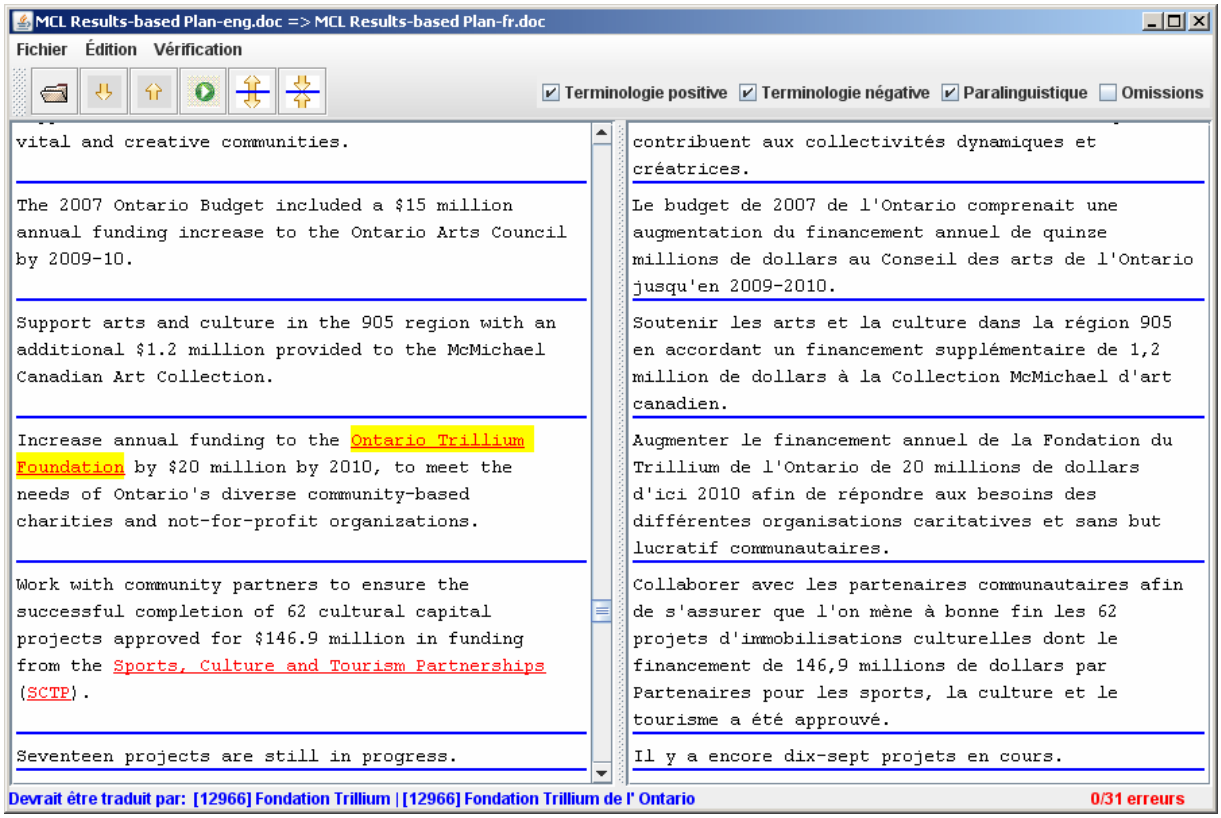


Figure 1: An example of TransCheck output

The source text appears on the left and the target text on the right; the dark horizontal lines mark the alignments that TC has automatically calculated. The term 'Ontario Trillium Foundation' is highlighted here because TC has not found in the aligned target segment one of the desired French terms, which are indicated on the bottom of the window.

results are output in a graphic user interface for the user to review. Figure 1 provides an example of TC output illustrating a potential error of positive terminology.

Now suppose that the user does not specify a target file when asking the system to check the positive terminology. TC will verify every source term that appears in its glossary and, not finding any target equivalents, will systematically flag each one. Odd as this may at first appear, it could in fact be very useful, especially when the output is saved in HTML format, since this HTML file can then be sent to outside translation suppliers, informing them of the terminology that the client requires in their translation. See Figure 2 below for an example of such output.

TC recently underwent an extensive field trial at the Government Translation Service (GTS) of Ontario, where the great majority of texts are outsourced to freelancers and other service providers who are contractually obliged to respect the terminology in ONTERM, the Ontario government's online terminology database. For various administrative reasons, GTS did not send its suppliers HTML files like that in Figure 2. However, the terminologists at GTS did supply them with a list of new terms not yet in ONTERM, which they extracted from the source texts using a specialized term-spotting module that the RALI had added to TC at their request. We will describe this new module in some detail below. First, however, we turn to a brief description of ONTERM, which provided the terminology that was used in the TC field trial at GTS.

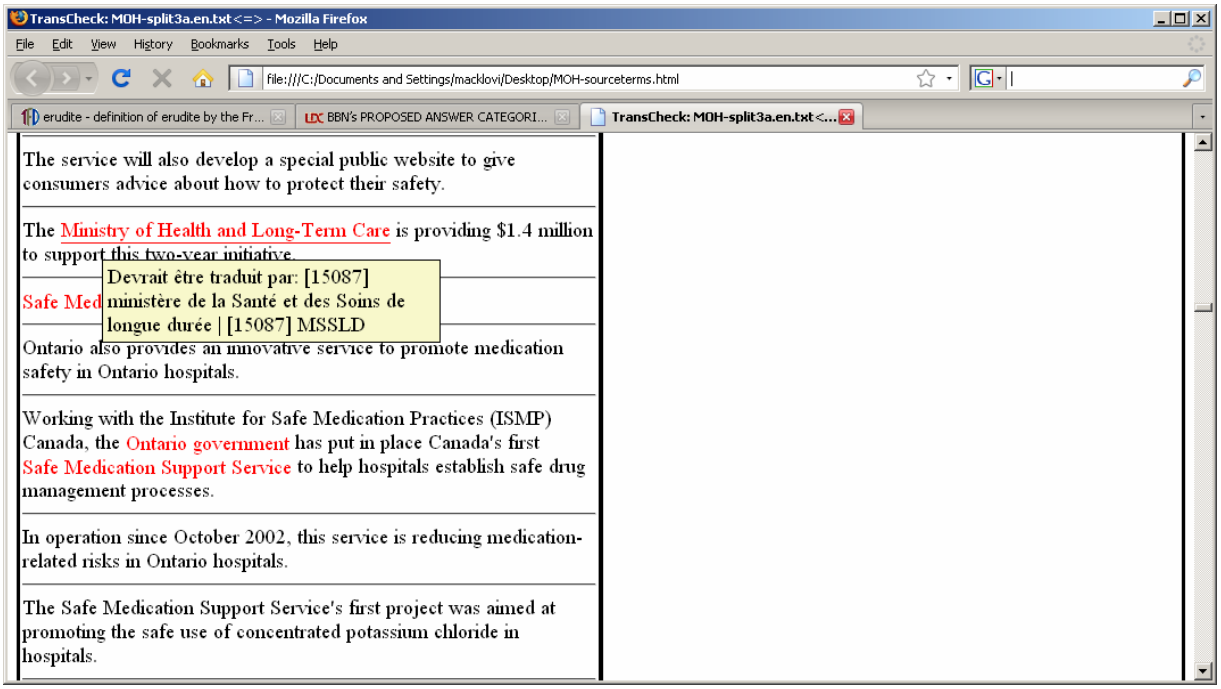


Figure 2: An example of monolingual output in HTML format

The user has previously specified a source text file and a glossary, and asked the system to verify the positive terminology. TC responds by highlighting all the terms that are present in the source text. The user saves the output in HTML format, which can then be sent to outside translation suppliers and viewed in any browser.

2 ONTERM

The demand for French translation within the Ontario government has been increasing steadily since 1986, when the government passed the French Services Act. In 1997, the Government Translation Service moved from an in-house translation model to the current outsource model which farms out almost all translations to private-sector suppliers. An in-house team of corporate translation advisors (CTAs) are responsible for managing a roster of suppliers and overseeing quality assessment and control. The Terminology Unit within GTS was given the mandate to create French equivalents for official Ontario Government English names, to ensure the consistent use of official names across Ontario Public Service (OPS) documents and to make names and terminology available to the OPS, translation suppliers and the public at large. To assist in accomplishing this mandate, the ONTERM database and website were created.[3] The ONTERM database focuses on official Ontario government names which include: the names of Ministries, agencies and organizational units; the names of councils, committees, working groups; the names of plans, programs, projects, policies and strategies; position titles; the names of IT systems and applications, conferences, exhibits, commemorative events, awards, distinctions, scholarships, catch phrases and political geographic entities.[4] ONTERM currently contains over twenty-six thousand terminology records.

[3] http:// www.onterm.gov.on.ca

[4] It will be noted that all of these, with the exception of the catch phrases, correspond to a particular set of *named entities*, whose defining characteristic is that the entity must fall within the jurisdiction of the Ontario government. For further discussion, see note 7 below, as well as Section 5.

3 Term checking in TransCheck

Term checking was a central component in the field trial of TC at GTS. It therefore behooves us to describe in somewhat more detail precisely how this operation proceeds in TC, beginning with the format of the entries in the system's terminological glossary.

3.1 Format of the entries in TransCheck's glossary

The following is a typical example of an entry in TC's terminological glossary:

```
(i) EN: [1619 ; 1619] Government Translation Service ; GTS
    FR: [1619 ; 1619] Service de traduction du gouvernement ; STG
```

The terminological equivalents that are to be verified by TC must be listed in a plain text file which the user loads at run time, and all the entries in the file must conform to this simple format: two fields, one for the English term, the other for the French term, with alternate terms (i.e. synonyms, abbreviations or other shortened forms) being separated by a semicolon. The system interprets this entry as follows: every time an occurrence of 'Government Translation Service' or 'GTS' is encountered in a segment of an English text, either 'Service de traduction du gouvernement' or 'STG' must be found in the aligned French segment; otherwise, a potential terminological inconsistency will be flagged. Note that such entries are bidirectional, i.e. they can be applied equally well to a French source text and an English translation, although the errors will always be flagged in the source. As for the numbers in square brackets, they refer to the original record in ONTERM and were inserted to help GTS' terminologists interpret TC's output; the system itself does not use them.

When TransCheck was originally conceived, the developers assumed (somewhat naively) that the users of the system would manually create the term glossaries, composed of entries in the above format, which the system would apply to draft translations. While this is certainly possible, it has the undesirable consequence of limiting the scope of the terminology that TC can verify; for few users are likely to manually key in thousands of entries like the one above. This is one reason why we reacted so favorably when we were approached by GTS regarding a possible field trial of TC; the terminologists there wanted to import into TC all the records in ONTERM and use the resulting system to help them vet outsourced translations.

Now obviously, many of the records in a bona fide term database will be far more complex than the one shown in (i) above, containing fields for the subject domain, definitions, observations and other notes, author and date of the record, etc.; and the same is true (though perhaps to a somewhat lesser extent) for the smaller in-house glossaries that are maintained by translation agencies and individual translators. Because we only import into TC the term equivalents from such records, some of the information on the records will inevitably be lost. In some cases, this will not be overly serious; in others, however, the information that is discarded may be crucial in helping a human translator decide if, or in what context, a particular term equivalence applies. Consider in this regard the following (non-hypothetical) example:

```
(ii) EN: aim
     FR: objet
      NOTE: Title of paragraph found in Course Management Details. It
          gives the main objectives of a course.
```

For a human translator, the note in (ii) is highly informative: it tells him/her in precisely what context the English term 'aim' can and should be translated as 'objet' in French. This information will be lost, however, when this record is imported into TC; and even if it were retained, the system would have no way of determining whether the text it was currently processing corresponded to a course lesson. Consequently, every time it encounters the word 'aim' in an English text, TC will seek the word 'objet' in the aligned French segment; and often it will not find it, because the word 'aim' allows for many other fully acceptable French equivalents, e.g. 'but', 'objectif', 'cible', and 'mire', to cite just a few.

3.2 Term invariability

As we saw in section 2, the majority of the records in ONTERM correspond to official government names. An important property of such appellations is that they tend to be invariant, in two senses; first, their form is relatively frozen, in not admitting the insertion of modifiers, for example; and second, these terms and their translation are unlikely to vary according to the linguistic context in which they appear. Consider the following two entries, taken from ONTERM and provided here in their simplified TC format:

```
(iii)  EN:[588] Chief Election Officer
       FR:[588 ; 588] directeur général des élections ; directrice géné-
           rale des élections

 (iv)  EN:[2882 ; 21329] Regional Director ; Area Director
       FR:[2882,21329 ; 2882,21329] directeur régional ; directrice ré-
           gionale
```

As we can see from the entry in (iii), it is not quite true to say that official names never vary, for the French designation for this position does change, according to whether the incumbent is a man or a woman. This being said, it is quite improbable that this multi-word term could accept other inflections, or admit an interposed modifier in either language. Notice too that the first letter of each English word in the term is capitalized, indicating that we are dealing here with the name of a particular official position. Hence, we are not likely to find the head noun of this term in the plural, because, like all proper nouns, it has a unique reference.[5] Synonymy with such terms is usually limited to abbreviations, acronyms or other shortened forms of the full term, as in the example in (i) above.

The entry in (iv) is slightly more complex; one way of interpreting it is as follows: the French term 'directeur regional' (or its inflectional variant 'directrice régionale') may be translated either as 'Regional Director' or as 'Area Director' – these last being two distinct terms. Notice as well that two different record numbers are included in the square brackets of entry (iv), indicating that the first English term comes from record #2882, while the second comes from record #21329. When we consult these records in ONTERM, what we find is that 'Regional Director' is the generally recommended term in the Ontario Public Service, while 'Area Director' is the term that is preferred in the ministry of Municipal Affairs and Housing. In other words, these two English terms name two distinct entities (which is why they have two distinct records), although both happen to use the same name in French. So here too we need to qualify the claim made above that official names tend to be invariant. The English designation for 'directeur régional' does indeed vary, although the conditioning factor is not *linguistic* but rather the organization to which the position belongs. The only solution open to TC in cases like this is to relax the condition used to flag terminological inconsistencies. Applying the entry in (iv) to a pair of aligned texts, TC will not flag a potential error upon encountering the French term if either 'Regional Director' or 'Area Director' is found on the English side; and this, even though one of those English terms may perhaps be incorrect in a text coming from a particular ministry. Because government organizational structures and position titles are often named the same way in one or other language in the various ministries across the Ontario government, though not always translated the same way, the TC ONTERM glossary contains a relatively high number of such entries: 1225 to be exact, or approximately 5% of the total entries, although many of the alternate terms turn out to be minor inflectional or orthographic variants, e.g. 'Board of Negotiation ; board of negotiation' or 'Pesticides Residue Section ; Pesticide Residue Section'.

---

[5] See (Quirk et al., 1972) for a succinct description of the key properties of proper nouns. Regarding the invariability of the English term in (iii), we do occasionally find instances in which the internal noun 'election' is pluralized: 'Chief Elections Officer'. While this may be correct in other jurisdictions or in other countries, ONTERM tells us that the correct form of the term, in Ontario at least, always has this noun in the singular.

3.3 Glossaries for human and machine use

There are a number of important lessons for automatic term checking that can be drawn from this discussion. The first, and perhaps the most general, is that the requirements of a term glossary intended for humans and one designed for automatic term checking are quite different. For the former, the comprehensiveness and detail of the information provided on each record are important attributes, because terminologists can rightly assume that humans are capable of intelligently interpreting such information. For a program like TC, on the other hand, the non-ambiguity of the entries is primordial, and this, for similar though inverse reasons. The program simply does not have the requisite intelligence to correctly interpret such information – neither the linguistic intelligence that would allow it to infer the grammatical dependencies that are often specified in the observations, and certainly not the real-world knowledge upon which humans instinctively rely to make the necessary distinctions that allow them to select the correct term. Entries like that in (ii) above are therefore anathema for TC. In general, we should not ask the system to verify the translation of vague or polysemous words that allow for multiple, equally correct target language equivalents. The system will fare far better when the terms it is asked to verify are linguistically complex (i.e. multi-word expressions), technical and highly specific, because such terms and their translations tend to remain invariable across different contexts.

Another lesson that the field trial at GTS served to underscore was obliquely alluded to above when we mentioned that TC was used there to help vet outsourced translations. Again, when TC was first conceived, the idea was that it would be primarily used to assist revisers. However, the manner in which the system was employed at GTS was more akin to quality control than to revision. The distinction between the two activities may not at first be apparent, but is in fact fundamental. In revision, a (normally senior) translator seeks to improve a draft translation by making changes to it.[6] The objective in quality control, on the other hand, is simply to determine whether or not a translated text meets certain quality standards; if it does not, it is refused or returned to the translator, without the quality controller necessarily having to make any changes to it. In its current state, our TC prototype lends itself better to quality control than to revision. The principal reason for this is that, while the user of TC can make changes to the target text and save those changes, the resulting modified text may not accurately replicate the format of the original source text, especially if the latter contains figures, tables or other elaborate layout.

4 Term-spotting with TransCheck

In addition to managing the Ontario government's translation needs, GTS is also responsible for maintaining ONTERM, the provincial government's online terminology database. A substantial part of this work involves keeping ONTERM up-to-date by adding to it new terms and their official equivalents. In principle, there are two ways in which GTS terminologists might go about this: they could either scour government documents on their own, with a view to identifying new terms that need to be added to ONTERM; or new terms could be brought to their attention by outside translation suppliers, who request that they provide an official French equivalent. Neither of these procedures is ideal: the first is labour-intensive and time consuming; the second is essentially reactive, whereas GTS would much prefer to be proactive.

In the course of the TC field trial, the terminologists at GTS asked the RALI if it could help them with this task by developing a specialized term-spotting module that would be integrated within TransCheck. Recall that a majority of the entries in ONTERM are official names that designate government positions, administrative units, programs, legislation, etc.; as such, most of these are proper nouns that begin in English with a capital letter. The suggestion of the lead terminologist at GTS was that the new term-spotting module in TC use this simple formal property as a diagnostic for identifying official names in English texts. Furthermore, since TC already incorporates a glossary containing all the terms in ONTERM, the new module should be able to determine which of the potential terms identified in a given text do not yet have entries in the database and hence might need to be added.

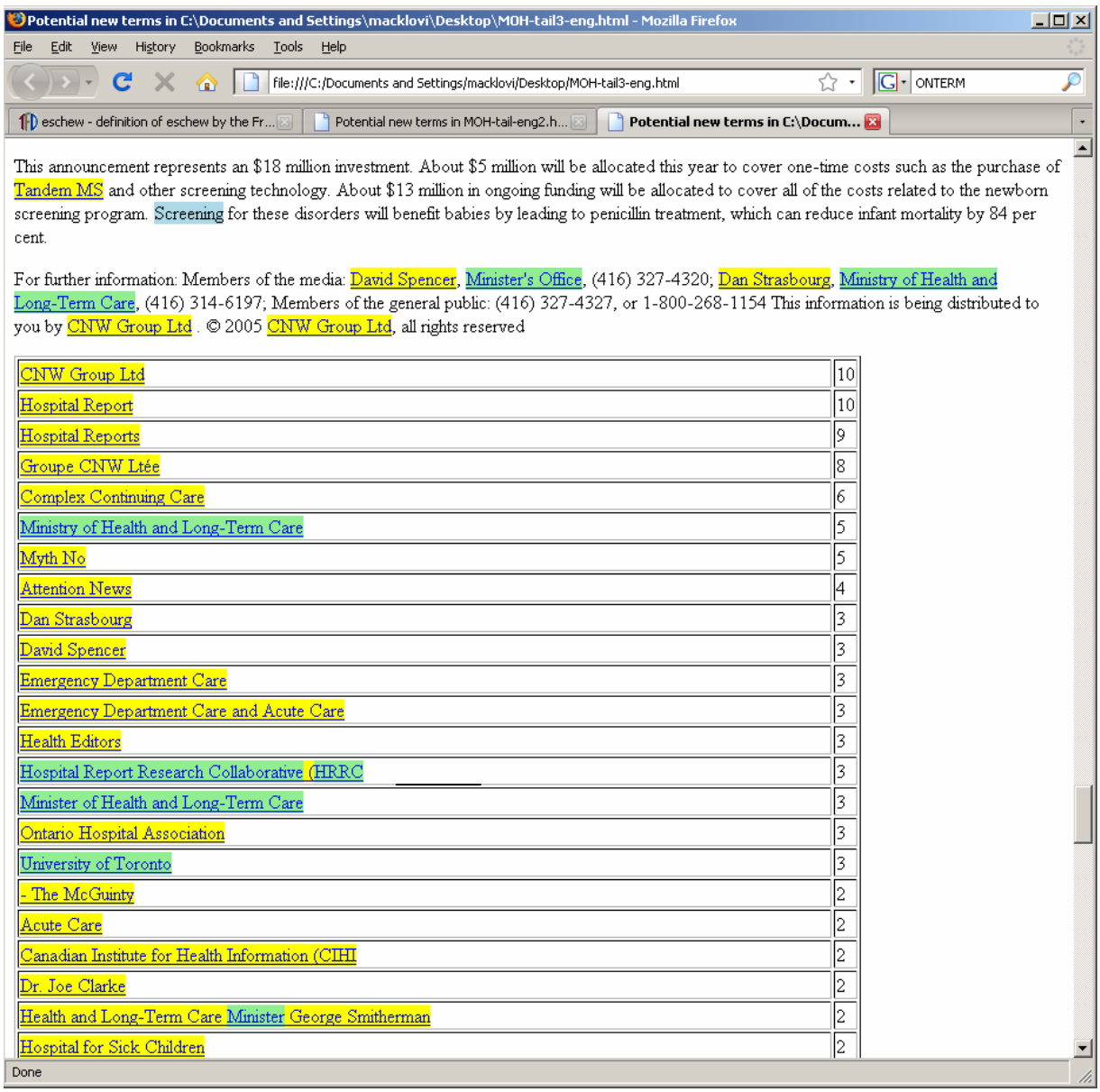
[6] Of course, a preliminary translation may also be reviewed and improved by the same translator who drafted it.

The RALI agreed to implement this suggestion in a new term-spotting module that was added to TC and subsequently tested at GTS. As before, TC begins by identifying all the terms in the English source text that are present in ONTERM. In a second pass, the new module then scans the text and locates all sequences of two or more words that begin with a capital letter. (Isolated, single words that begin with a capital are simply too ambiguous; this would result in the first word of every sentence being proposed as a potential term and would significantly increase the program's noise level.) The capitalized words in these sequences are generally contiguous, although the program does allow for a small number of lower-case 'skip words', in the form of certain common prepositions (e.g. 'of' and 'for'), the articles 'the' and 'a', and the conjunction 'and'. When the entire text has been processed in this way, TC outputs an HTML version that is colour-coded in the following way: terms already present in ONTERM are highlighted in blue; potential new terms are highlighted in yellow; those multi-word expressions that begin with a capital and already have an entry in ONTERM are highlighted by both routines and so appear in green. TC also produces a table at the end of this HTML file listing all the potential new terms and their frequency. Entries in the table are hyperlinked to their occurrences in the text, so that a user can easily inspect the candidate terms in their context. An example of this HTML output is given in Figure 3 below.

A quick glance through this list shows that many of the proposed multi-word sequences are not in fact true terms and hence should not be added to ONTERM. Among them are proper names, either of persons (e.g. 'Dan Strasbourg' or 'Dr. Joe Clarke') or of companies (e.g. 'CNW Group Ltd') which do not belong in a term bank. Other dubious entries in the table arise from problems in properly segmenting the source text. For example, at the beginning of each of the press releases that make up this text, there is a call to a particular set of addressees: 'Attention News/Health Editors'. Our program does not properly parse the scope of the slash as a conjunction and so this sequence is mistakenly segmented into two potential terms, the first of which is incoherent. A similar problem occurs with the period at the end of 'Myth No', which should normally include the following number, e.g. 'Myth No. 5'. (Whether this should be considered a true term is another question.) And then there are instances of noise which are simply unavoidable, given the simplicity of our term-spotting criteria. For example, 'The McGuinty' comes from the beginning of a sentence that continues '…government is encouraging Ontarians to protect themselves'. Here we have a sequence of two words that begin with a capital letter, although this is certainly no term.

On the other hand, this short portion of the table shown in Figure 3 also contains a fair number of entries which do seem to constitute bona fide terms, and hence might have to be added to ONTERM: 'Complex Continuing Care', 'Emergency Department Care', 'Ontario Hospital Association', 'Canadian Institute for Health Information', and 'Hospital for Sick Children'.[7] The status of other entries, e.g. 'Hospital Reports' is perhaps less clear and would require the considered judgment of a qualified terminologist who would begin by verifying the occurrences of this expression within the text – something s/he could easily do by means of the inserted hyperlinks. The important point is this: despite the noise that such lists may contain, qualified terminologists have no trouble running through them and rapidly picking out new terms for inclusion in the database.

[7] In point of fact, the last three terms would not be added to ONTERM, because they either designate bodies that are outside the Ontario government's jurisdiction (the *Canadian* Institute for Health Information), or because they are not actually part of the government itself (both terms involving hospitals). This is a good illustration of the subtly of the knowledge that is required to distinguish between terms that should or should not be added to ONTERM.



Potential new terms in C:\Documents and Settings\macklovi\Desktop\MOH-tail3-eng.html - Mozilla Firefox

This announcement represents an $18 million investment. About $5 million will be allocated this year to cover one-time costs such as the purchase of Tandem MS and other screening technology. About $13 million in ongoing funding will be allocated to cover all of the costs related to the newborn screening program. Screening for these disorders will benefit babies by leading to penicillin treatment, which can reduce infant mortality by 84 per cent.

For further information: Members of the media: David Spencer, Minister's Office, (416) 327-4320; Dan Strasbourg, Ministry of Health and Long-Term Care, (416) 314-6197; Members of the general public: (416) 327-4327, or 1-800-268-1154 This information is being distributed to you by CNW Group Ltd . © 2005 CNW Group Ltd, all rights reserved

CNW Group Ltd	10
Hospital Report	10
Hospital Reports	9
Groupe CNW Ltée	8
Complex Continuing Care	6
Ministry of Health and Long-Term Care	5
Myth No	5
Attention News	4
Dan Strasbourg	3
David Spencer	3
Emergency Department Care	3
Emergency Department Care and Acute Care	3
Health Editors	3
Hospital Report Research Collaborative (HRRC	3
Minister of Health and Long-Term Care	3
Ontario Hospital Association	3
University of Toronto	3
- The McGuinty	2
Acute Care	2
Canadian Institute for Health Information (CIHI	2
Dr. Joe Clarke	2
Health and Long-Term Care Minister George Smitherman	2
Hospital for Sick Children	2

Figure 3: Output of TransCheck's term-spotting module

The table lists all multi-word sequences that begin with a capital letter. Those highlighted in yellow are not found in ONTERM and so correspond to potential new terms; their frequency is given in the right-hand column of the table. Those terms that are found in ONTERM are highlighted in blue; we see one occurrence of 'screening' in the text above the table. Capitalized multi-word sequences that are identified by our term-spotting module and also appear in ONTERM receive both blue and yellow highlighting, and so they appear as green. Each entry in the table is hyperlinked, allowing the terminologist to quickly peruse all its occurrences in the text.

5 Discussion

Automatic term identification and extraction – what we have been calling term-spotting – has a relatively long history, going back at least to the seminal paper of (Justeson & Katz, 1993)[8]. What these researchers demonstrated is that a large proportion of the technical terminology in English texts corresponds to nominal groups (i.e. nouns and their pre- and post-modifiers) that can be accurately identified using regular expressions defined over part-of-speech categories. Moreover, a simple but surprisingly effective way of distinguishing technical terms from ordinary noun phrases is to use the criterion of full repetition. Multi-word technical terms tend to reappear verbatim in a text, whereas non-technical, descriptive noun phrases do not; upon repetition, they are often pronominalized or truncated in various ways. Subsequent research has elaborated on this basic approach. (Daille, 1994), for example, extended it to French and showed that lengthy complex terms could profitably be analysed as combinations of simpler term sequences that are generated by means of various devices such as co-ordination. In an effort to increase the precision of the extracted candidate terms, (Drouin, 2003) compares the frequencies of nouns and adjectives in a technical corpus under analysis with their frequency in a non-technical, general corpus of the language – the idea being to extract only those terms made up lexical items that are highly specific to the technical domain. The results are quite encouraging, particularly for single-word terms, which are often ignored by term extraction programs because their recognition is problematic. (Patry & Langlais, 2005) propose an alternative to a handcrafted static definition of what constitutes a legitimate term, specified as possible sequences of POS tags; from a corpus of term examples supplied by the user, they train a language model that automatically generates these POS patterns.

Compared to these approaches, the term-spotting algorithm that we have implemented in TransCheck and described above appears extremely simple, not to say simplistic. We do no part-of-speech tagging, do not calculate any statistical correlate of term likelihood (e.g. mutual information), and do not compare the frequency of the components of our candidate terms with their frequency in a non-technical reference corpus. Instead, we define one simple pattern for our candidate terms, based on whether their component words begin with a capital letter. The reason we can do this, of course, is that the terms we are looking for are all official names, and official names are proper nouns which all begin with a capital letter – at least in English.[9] The problem, however, is that there exist other types of proper nouns which do not designate official government appellations, e.g. personal names, temporal names and geographical or place names. These too begin with a capital letter and are the source of much of the noise in the table of candidates terms that TC produces.[10]

Now in principle, it would be possible to automatically filter out some of this noise by incorporating within TC additional linguistic machinery. Consider personal names, for example, several of which appear in the list reproduced in Figure 3 above. There has been much work on named entity recognition in recent years, and many programs now exist which can reliably identify personal names in running text. The problem is that ONTERM contains many records for terms that include personal names, e.g. 'Sir John A. Macdonald Highway', 'Philip Shrive Memorial Bridge', 'Lincoln M. Alexander Building', Dr. Albert Rose Bursary'. This probably explains the less-than-enthusiastic response of GTS terminologists to our suggestion of adding such a filtering component to TC. Given their ability to rapidly scan the candidate terms and accurately pick out those that should be added to ONTERM, they prefer to retain the noise rather than run the risk of missing a potential new term. And so we have left TC as is, imperfect in theory perhaps, but quite adequate in practice.

---

[8] Although this paper was published in 1993, it was actually written and disseminated several years earlier.

[9] Hence, our term-spotting algorithm wouldn't work with source texts in French, where proper nouns do not follow the same capitalization rules as in English, or in German, where all nouns are capitalized. But at GTS, the overwhelming majority of source texts are in English.

[10] Actually, some geographical names are included in ONTERM, e.g. those (like 'Thousand Islands Parkway') that serve to designate government buildings, highways, historical sites, etc. Moreover, the ONTERM Web site provides access to the GeoNames database, which lists over fifty-seven thousand geographic names in English and French.

6 Conclusion

A recognized feature of good writing and good translation, especially in electronic environments, is terminological consistency. Urgent texts frequently have to be divided up among several translators or outside translation suppliers; the reviser – or in the case of GTS, the corporate translation advisor – is then left with the task of merging these parts into a coherent whole. Ensuring terminological consistency is an important part of this job, and it can be quite onerous. The trial at GTS has shown that TransCheck can help by automating this task to a considerable extent. Indeed, CTAs found that simply by aligning the source and target texts in side-by-side format, TC allowed them to perform quality control on the entire text (not just the terminology) more efficiently and thoroughly than in the past.

For the Ontario government, this question of terminological consistency is particularly important. In today's electronic environment, the consistent use of correct terms, and especially names, is essential for accessing information. We all know how helpful it is, in conducting a successful Web search, to be able to query the correct designation of what one is looking for. But terminological consistency is also important for developing a brand and a brand personality. While in the private sector, the focus is mainly on company and product names, within a government context, the official names used across ministry websites play a special role in projecting the government's brand personality and fostering a relationship with its citizenry. Since ministry home pages increasingly act as the electronic doorways to virtual government offices, it is crucial that clear and consistent names in both English and French, and between English and French, be used to project the government's personality, to find information and to navigate effectively among its numerous websites and pages.

The latest version of TransCheck with its term spotting module has allowed the Terminology Unit in GTS to make more efficient use of its research time. During the period of the field trial, the Terminology Unit handled a 53% increase in the number of terms rendered into French, in large part on account of the use of the new TC system to pre-process the government's results-based plans. Not only does TC give the terminology service the ability to be more systematic and thorough in the detection of new Ontario government terminology, it allows it to be more proactive, thereby reducing the amount of after-the-fact checking. Making corrections after translations have been submitted is both costly and time-consuming. Sometimes, they are so costly that the corrections are never made. Being proactive promotes the philosophy of 'doing it right the first time'.

References

Daille, Béatrice. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* PhD thesis, Paris : Université de Paris 7.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99-115.

Justeson, John & Slava Katz. 1993. Technical terminology: some linguistic properties and an algorithm for identification in text. Technical Report RC 18906, IBM Research Division.

Macklovitch, Elliott. 1994. Using Bi-textual Alignment for Translation Validation: the TransCheck System. *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, pp.157-168.

Patry, Alexandre & Philippe Langlais. 2005. Corpus-Based Terminology Extraction. *7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Danemark, pp. 313-321.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1972. *A Grammar of Contemporary English.* Longman Group Ltd., London.

Linguists to engineers:

Explaining the meaning of *service*

Heli Tissari
P.O. Box 24 (Unioninkatu 40 B)
FIN-00014 University of Helsinki
Finland
heli.tissari@helsinki.fi

Tuula Chezek
P.O. Box 24 (Unioninkatu 40 B)
FIN-00014 University of Helsinki
Finland
tuula_jack@hotmail.com

Abstract

This paper was originally written for engineers at a technological research centre who are doing service science.[1] It is a pilot study which we hope to develop in the future. Here, we outline the history of the noun *service* as presented by the *Oxford English Dictionary* (OED), some electronic corpora of the English language, and the Internet. The corpora which we used represent eighteenth-, nineteenth- and twentieth-century English. The OED suggests that the medieval 'waiting at table, supply of food' sense of *service* is the origin of all senses suggesting the 'supply of something', such as the present-day English 'provision of maintenance or repair work to ensure the efficient running of something'. In the nineteenth century, we see such modern senses of service as 'the supply of gas, water etc. through pipes from a reservoir', 'provision of labour and materials for the carrying out of some work for which there is a constant public demand', and 'transit afforded by vehicles plying regularly on a route'. In this way, the history of the word reflects the development not only of modern society, but of modern technology. The gist of this paper lies not only in its topic but also in its somewhat a-theoretical form: It is the end product of a series of discussions with representatives of another field of research.

1 Introducing the noun service

1.1 *Service* is a very productive noun

One way to look at nouns is to observe the amount and variety of compounds they produce. Our investigation showed *service* to be quite a productive noun.[2] Above all, it tends to qualify the so-called head noun, such as *agreement* or *manual*, exemplified in the following (read *service* for *s*):

[1] They commissioned it in order to publish it elsewhere, but gave us the permission to present the same findings at the International Workshop on Terminology and Lexical Semantics (TLS'09). The title of the original article is "Animals, goods and technological innovations: The story of the noun *service*". This is a slightly modified version. The original purpose of the article is nevertheless the reason why it includes few references. For theoretical background, see e.g. Tissari 2003, which includes a lengthy introduction to the study of meaning.

[2] We used the *WebCorp* service for linguists and AltaVista search engine. <http://www.webcorp.org.uk/> was accessed on 9 September 2008 and produced 1,147 instances of the noun *service* from 200 web pages, 29 of which returned errors.

s access point, s account, s agreement, s center, s contract, s creation, s delivery, s department, s desk (team), s dispatch, s encounter, s experience, s guarantee, s industry, s information, s interaction, s job, s-learning, s manager, s manual, s menu, s number, s object, s opportunity, s option, s order, s organization, s output, s pack(age), s personnel, s process, s product, s profession(al), s program, s project, (Internet) s provider, s provision, s quality, (military) s records, s request, s reseller, s road, s security, s software, s solution, s station, s supplier, s support, s time, s volume, s work(er)

However, service can also occur as the head itself, as in:

anesthesiology s, bus s, car s, church / divine / Sunday s, community s, customer s (department / operation), energy s, funeral s, lip s, tea s, telegraph s, telephone s, television repair s, translation s, (satellite) TV s, voice s, wire s.

1.2 *Service* as used in *service science*

In the compound *service science*, *service* qualifies the head *science*. The focus falls on the latter word. According to the *Oxford English Dictionary* (OED), these kinds of business-related compounds were not frequent in the English language until the twentieth century, although the compound *service-side* was already being used in the eighteenth century in the context of tennis.[3] Twentieth-century examples from the OED include *service industry*, *service occupation*, and *service sector*. It is likely that these compounds have proliferated, especially towards the latter end of the twentieth century. This is suggested by a word-search for *service* in the one-hundred-million-word *British National Corpus* (BNC), which reveals this word to be twice as frequent in British English texts published in 1985–1993 as in British English texts published in 1960–1974.[4]

Interestingly, *service science* is spelled in several ways. For example, the Wikipedia entry for "Service Science, Management, and Engineering" has *service science*, *services sciences*, and *a science of service*.[5] Chesborough and Spohrer use *services science* (2006). It seems that we are still dealing with an emerging concept with no fixed name.

Service science provides the following definition for the concept of 'service': "A service is a change in the condition of a person, or a good belonging to some economic entity, brought about as the result of the activity of some other economic entity, with the approval of the first person or economic entity." (Hill quoted by Chesborough & Spohrer 2006: 36.) We may note that many of the compounds presented above reflect this definition. They deal with the people providing the services, the kind of services provided, and with the process of supplying the services.

2 Outlining the history of the noun *service*

In this section, we will outline the history of the noun service as presented by the OED and some electronic corpora of the English language. The corpora which we have used for this section represent eighteenth-, nineteenth- and twentieth-century English. This is a pilot study, meaning that the corpus results are tentative.

---

[3] We have used the *Oxford English Dictionary Online* service, <http://dictionary.oed.com>, accessed 2 June 2008.
[4] This query does not differentiate between the various senses of *service*, and does not include the plural form *services*. We queried in the BNCweb CQP-edition on 18 June 2008, and thank prof. Terttu Nevalainen for the tip. The frequencies per million words are 311.24 and 148.39, respectively.
[5] <http://en.wikipedia.org/wiki/Service_Science,_Management_and_Engineering> accessed 18 June 2008.

The following is a simplified presentation of the OED's main senses for the noun *service*:[6]

- The condition of being a servant to a master (I).
- What(ever) someone does in order to serve a superior (II).
- Religious devotion and/or activity (III).
- Something given or received (IV).
- Supply of something (mainly material, but potentially immaterial) (V).

In addition to these five senses, the OED lists a great number of subsenses, part of which will be discussed below. Table 1 shows the order in which these senses have been introduced to the English language, according to the OED. Notably, each of these senses has existed since medieval times. Both the noun *service* and the verb *to serve* can be traced back to Old French and, eventually, to Latin, where *servus* meant 'servant'.

Table 1. Introduction of the main senses of *service* to the English language.

Sense	First attestation
I. The condition of being a servant to a master.	1250
II. What(ever) someone does in order to serve a superior.	1200
III. Religious devotion and/or activity.	1100 (tentative)
IV. Something given or received.	1325
V. Supply of something (mainly material, but potentially immaterial).	1300

2.1 The Middle Ages (ca. 1100–1500)

All the main elements of the present-day definition of *service* presented in the introduction (Hill quoted by Chesborough & Spohrer 2006: 36) were already present in the meaning of *service* in the Middle Ages. OED's first sense, 'the condition of being a servant to a master' suggests that we are dealing with specific persons and their respective conditions, and that this is likely to involve some kind of contract between these people. The exchange of goods or commodities is particularly well illustrated by the OED subsense 'that which is served up or placed on the table for a meal'.

The OED in fact suggests that the 'waiting at table, supply of food' sense of *service* is the origin of all senses suggesting the 'supply of something', such as the present-day English 'provision of maintenance or repair work to ensure the efficient running of something'. Here we can see a generalisation of the sense 'supply of food' to 'supply of anything that people need'. Think of the medieval church as well, with its wealth in buildings and ritual objects, on top of rules regarding worship. The noun *service* in medieval English meant both the 'rules' and the 'celebration' of the ritual. A comparison of these with present-day service protocols and their enactments may not be trivial. We invest in rules where they are important for us.

[6] We are now concerned with the first entry for *service* in the OED. The second entry concerns a completely different meaning, a type of tree.

2.2 The Early Modern period (ca. 1500–1700)

The OED sense 'duty of a soldier or sailor', including the contract, the moral obligation, and the actual performance of the duty, stands out as a potential core sense of *service* in the sixteenth century. The OED quotes Shakespeare's *Much Ado About Nothing*, where a character says of another that

(1) He hath done good *seruice* … in these wars. (1599)

This may be generalisation from the sense 'being servant to a master' to the sense 'being servant to a superior', such as a military commander, or even a nation.

The other 'duty' sense of *service* in the Early Modern period is somewhat different. It is exemplified by a greeting from an Early Modern English letter:[7]

(2) Present my affectionate love and *service*, good Madam, to my good cosin … (Thomas Meautys 1626)

Service in this sense appears in polite expressions, suggesting that the writer or speaker wishes to be friendly towards and stay on good terms with the person addressed. Here, the 'feudal allegiance' sense of *service* has become a sign of respect, rather than referring to a relationship between a slave and a master.

This nicely leads us to the sense 'conduct tending to the welfare or advantage of another'. Such behaviour may be motivated by a sense of duty, or a wish to get something in return, but it may also contain elements of spontaneity, voluntary will and altruism. These may be involved in the OED example

(3) I intend to do you *service* by revealing to you my very heart. (Patrick: *The Parable of a Pilgrim* 1663)

2.3 The eighteenth century

We searched for the noun *service* in eighteenth-century texts included in the *Corpus of Late Modern English Texts* (CLMET). Our findings suggest that in that period, *service* often involved instances of good will, or simple wishes to help someone, which might still very well be core components of *service* today (OED sense IV 'something given or received'). Furthermore, figure 1 suggests that the OED's second sense 'what(ever) someone does in order to serve a superior' is the most frequent in this period. Although *service* could mean 'the condition or employment of a public servant' even in medieval times, this and related senses only seem to proliferate later.

[7] This example comes from the *Corpus of Early English Correspondence*, an electronic resource containing early English letters. See <http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html> for more details.

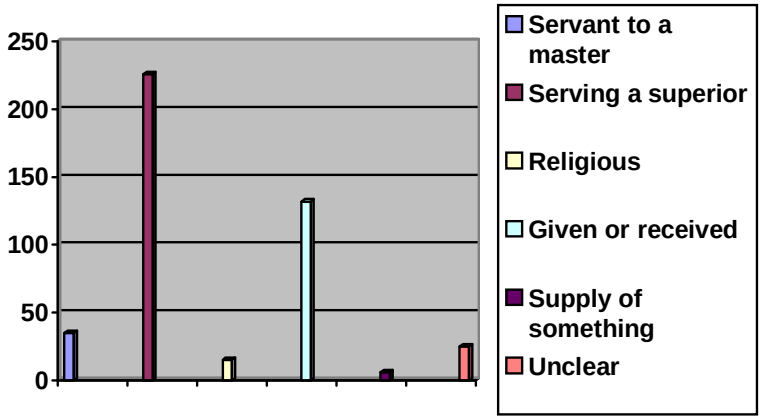


Figure 1. *Service* in the eighteenth century data (439 hits in ca. 2 mio words).

2.4 The nineteenth century

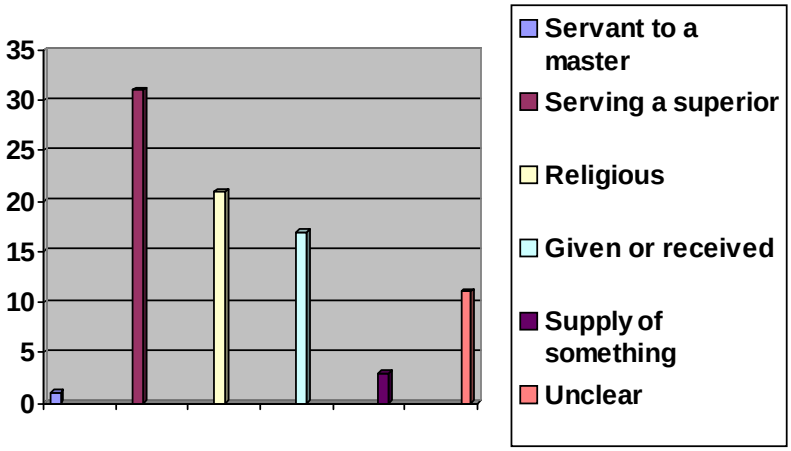


Figure 2. *Service* in the nineteenth century data (84 hits in ca. 362,000 words).

In the nineteenth century, we see such modern senses of *service* as 'the supply of gas, water etc. through pipes from a reservoir', 'provision of labour and materials for the carrying out of some work for which there is a constant public demand', and 'transit afforded by vehicles plying regularly on a route'. In this way, the history of the word reflects the development not only of modern society, but of modern technology.

Our nineteenth-century data from *A Representative Corpus of Historical English Registers* (ARCHER) suggest that 'serving a superior' was still the main sense of *service*, followed by the 'serving of God' (figure 2). It seems that in this period, the emphasis on a "giving heart" continued, although the 'business/society' sense of *service* also started to take root.

2.5 The twentieth century

The history of *service* infolds the kinds of things and the manners in which these can be supplied in each period. The 20th-century meaning of *service*, 'the serving of a meal in the restaurant-car of a railway train or on a ship' is only made possible by the invention of trains on the one hand, but also becomes relevant through more people being able to afford travel. *Service station* is a related case: these become relevant when masses of people begin to buy and travel in cars. (Cf. section 1.1.)

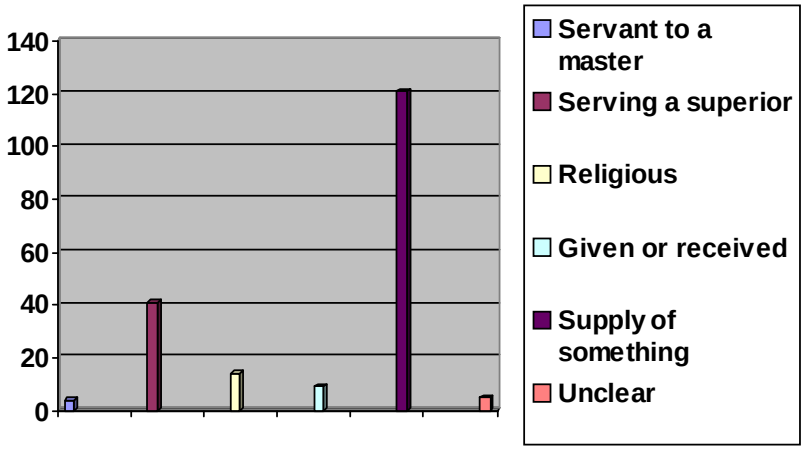


Figure 3. *Service* in the twentieth century data (193 hits in ca. 1 mio words).

The OED suggests a few novel senses of *service* in the 20th century. Another is 'expert advice or assistance given by manufacturers and dealers to secure satisfactory results from goods made or supplied by them', which occurs as early as 1919. A little later we get 'the supply of programmes by a particular broadcasting station' (1927), and then, 'the section of the economy that supplies needs of the consumer but produces no tangible goods' (1936). Later, this last sense is extended to compounds such as *service science*, where *service* means 'of or pertaining to services'.

We looked at the *Freiburg-Brown Corpus of American English* (FROWN) to see how *service* was used in 1991. Our findings suggest that the 'business/society' sense of *service* has grown into a giant plant leaving all the other senses in its shadow.

3 Discussion

3.1 A summary of the corpus findings

In terms of occurrences per 10,000 words, ARCHER (19th century) provides the most instances of the noun (2.3). CLMET (18th century) comes second (2.1) and FROWN (20th century) third (1.9). In this light, it seems that the use of the noun has been on the decline, which is contradicted by its recent productivity and by the results from the BNC (cf. section 1.2.).

Let us then look for the most frequent sense of each time period: Sense II 'what(ever) someone does in order to serve a superior' receives the most matches in CLMET (18th century) as well as in ARCHER (19th century). FROWN (20th century) on the contrary offers a totally different view, with sense V 'supply of something' being the most frequent (121 out of 193 = 63%). That is a drastic rise from only 6 (out of 439 = 1.4%) in CLMET and 3 (out of 84 = 3.6%) in ARCHER.[8]

3.2 Privatization

We have already used the expression 'business/society' sense of *service* to describe its development in the nineteenth and twentieth centuries. The two ideas merge in particular in the concept of privatization (state services being owned by private entrepreneurs). This development is reflected, for example, in the compounds *service-learning* and *community service leaders* and the concepts behind them. Consider example (4):

(4) Nokia and IYF are not alone in that belief. In fact, there are several other public and private organisations working to "harness the power" of youth and to create a new generation of *community service leaders*. Two examples are The Congressional Youth Leadership Council (CYLC) and the National Youth Leadership Council, both of which focus on conferences. The latter promotes "*service learning*" and hosts an annual gathering of high school students aimed at introducing them to social, political and environmental issues, empower them to take action and encourage them to co-operate with people from different racial, geographic and socio-economic backgrounds. (Alison Beard in *Financial Times* 23 January 2004)

3.3 The Internet

An even more recent development is the growth of services in and via the Internet, which contributes to new compounds with and without *service*. Consider (5):

(5) The atmosphere is being recreated in the tech world with a smoky haze hanging over the rolling concept of *cloud computing* – just a buzz phrase or perhaps the biggest thing to happen in computing since e-commerce? --- If cloud computing sounds a bit nebulous, here are a couple of definitions. Gartner defines it as "a style of computing where massively scalable IT-related capabilities are provided 'as a *service*', using internet technologies, to multiple external customers". (Chris Nuttall in *Financial Times* 8 July 2008)

Consider also table 2, which shows that on the Internet, compounds with *service* most often relate to the Internet itself. It becomes relevant to ask to what extent this reflects the current meaning of *service* and whether OED should revise its sense V 'supply of something (mainly material, but potentially immaterial)'. May Internet services be considered mainly material, or is the meaning of *service* changing in this respect?

[8] It is also interesting that sense III 'religious devotion and/or activity' is frequent in ARCHER but unfrequent both in CLMET and in FROWN. This may reflect an actual shift its popularity, but may also result from choices of texts in each corpus.

Table 2. Some Google results (accessed 1 August 2008).[9]

Compound	Frequency in millions	Ranking
civil service(s)	47.3 (34.4) > 81.7	9.
customer service(s)	309 (79.4) > 388.4	3.
phone service(s)	72.9 (51.7) > 124.6	6.
public service(s)	156 (198) > 354	4.
web service(s)	291 (293) > 584	1.
Yahoo service(s)	53.3 (337) > 390.3	2.
service(s) animal	19.6 (16.8) > 36.4	11.
service(s) directory	62.2 (60.1) > 122.3	7.
service learning (service-learning)	57.8 (2.77) > 60.57	10.
service public (service-public)	144 (7.6) > 151.6	5.
service quality	97.4 (4.28) > 101.68	8.

3.4 Potential associations in a person's mind

People's associations of the noun *service* with various concepts certainly depend on their personality and background, including native language. It is therefore impossible to say exactly what people think about when they use or hear the word, but we may venture an educated guess at this point. Historical strata of the meaning of *service* are likely to mix, for example as suggested in figure 4. In this imaginary mind, the original 'servant' sense of *service* connects through 'duty' and 'society' to 'privatization', while the other core sense, 'generosity', is associated with 'business' ('luxury') and finally with 'privatization'.

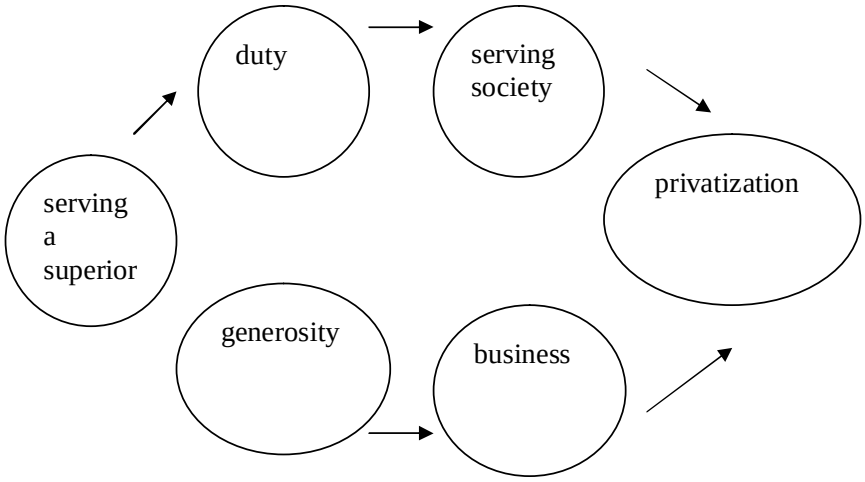


Figure 4. *Service*: potential associations.

[9] Note that we have googled for two different forms of a compound, with the plural in brackets. While the results indicate which form is preferred, in reality the searches do not completely distinguish between the singular and the plural. These searches, which were based on experimenting with the *WebCorp* service, also did not wholly distinguish between English and French, the latter being represented by *service(-)public*.

There seem to be two relevant questions. One is which associations are primary and which secondary; the other related question is whether people associate *service* with something positive or negative. The relative importance or 'weight' of the various associations contributes to how positive or negative a matter *service* is. The two concepts of 'generosity' and 'what a servant does to serve a master' are likely to contribute to a positive reading of *service* in business contexts, suggesting that the client enjoys some kind of privilege. If this is so, it may be difficult to advertize *self-service*, a compound in use since 1919 (OED).[10]

People's understanding of a word is of course rooted in their encounters with it. Consequently, the frequency in text and speech of such compounds as listed in sections 1.1. and 3.3. is also quite relevant to how people conceptualize *service*.

4 Conclusion

The service science definition of 'service' involves people, activities and economic entities (Hill quoted by Chesborough & Spohrer 2006: 36, cf. section 1.2.). We have shown how the noun *service* has, and may still, refer to various kinds of economic exchanges between people, and how its meaning reflects developments both in society and technology. We have also shown that Hill's definition misses the notion of 'generosity', even 'voluntary kindness', which may be quite important as well. Let us finally mention services provided by animals, closing with a final example of a compound:

(6) *Service animals* are animals that are individually trained to perform tasks for people with disabilities such as guiding people who are blind, alerting people who are deaf, pulling wheelchairs, alerting and protecting a person who is having a seizure, or performing other special tasks. *Service animals* are working animals, not pets.
<http://www.ada.gov/svcanimb.htm> accessed 30 July 2008

References

Chesborough, Henry, & Jim Spohrer. 2006. A research manifesto for services science. *Communications of the ACM*, 7(49:) 35–40.

Tissari, Heli. 2003. Love*scapes: Changes in prototypical senses and cognitive metaphors since 1500*. Mémoires de la Société Néophilologique de Helsinki LXII. Helsinki: Société Néophilologique.

[10] We thank prof. Matti Kokkala for bringing the compound *self-service* to our attention. We checked the BNC for it, but the corpus only provided 121 instances, which is little when compared to the 30,255 instances of *service*. We queried in the BNCweb CQP-edition on 10 September 2008.

Intégration d'informations syntaxico-sémantiques dans les bases de données terminologiques : méthodologie d'annotation et perspectives d'automatisation

Fadila Hadouche
Laboratoire RALI, DIRO
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
hadouchf@iro.umontreal.ca

Marie-Claude L'Homme
Observatoire de linguistique Sens-Texte
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
mc.lhomme@umontreal.ca

Guy Lapalme
Laboratoire RALI, DIRO
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
lapalme@iro.umontreal.ca

Annaïch Le Serrec
Observatoire de linguistique Sens-Texte
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
Annaich.le.serrec@umontreal.ca

Résumé

Dans le présent article, nous décrivons un modèle accompagné d'une méthodologie visant à expliciter les propriétés syntaxico-sémantiques de termes. L'explicitation est réalisée au moyen d'une annotation des propriétés de termes de nature prédicative et de leurs actants dans des contextes extraits de corpus spécialisés. Le projet comporte deux volets. Le premier consiste à définir le modèle à partir d'une annotation manuelle. Le second consiste à mettre au point une méthode d'annotation automatique. Nous décrivons les objectifs du projet ainsi que son l'état d'avancement.

1 Introduction

Les dictionnaires spécialisés et les bases de données terminologiques, quoique riches en information de nature conceptuelle, fournissent en général très peu de renseignements sur les propriétés linguistiques des termes ou sur leur comportement en langue. Quelques exceptions, toutefois, ont commencé à faire leur apparition : des dictionnaires spécialisés conçus dans une perspective d'apprentissage (par exemple, le Binon *et al.* 2000 dans le domaine des affaires) ou des bases terminologiques incorporant des rubriques sur le fonctionnement syntaxique ou sur la combinatoire des termes (par exemple, Lépinette & Olmo Cazevieille 2007; DiCoInfo dans le domaine de l'informatique). En outre, les bases de données lexicales sont de plus en plus sollicitées dans de nombreuses applications en traitement des langues naturelles (TAL). Il devient donc nécessaire de fournir sur les unités lexicales – qu'elles soient générales ou spécialisées – une description explicite et formelle de l'ensemble de leurs propriétés linguistiques.

Le présent projet s'inscrit dans cette mouvance. Il vise à proposer un modèle ainsi qu'une méthodologie permettant d'intégrer aux bases de données terminologiques une explicitation des propriétés syntaxico-sémantiques des termes. Les propriétés ainsi décrites peuvent se prêter à trois applications : 1. offrir aux utilisateurs un accès à des renseignements liés au fonctionnement des termes dans les textes; 2. permettre aux terminologues d'appuyer leurs descriptions sur des réalisations véritables

en corpus et de confirmer ou d'infirmer leurs intuitions; 3. intégrer à des applications de TAL des descriptions qui rendent compte des propriétés syntaxiques et sémantiques des termes et des liens entre ces deux ensembles de propriétés.

L'explicitation est réalisée au moyen d'une annotation formelle des dites propriétés réalisés dans des contextes dans lesquels les termes apparaissent. Le projet comprend deux volets : 1. le premier volet consiste à mettre au point le modèle d'annotation à partir d'une analyse manuelle d'un nombre important de contextes; 2. le second volet vise à mettre au point une méthode d'annotation automatique dont l'amorce repose sur un apprentissage réalisé à partir des contextes traités manuellement.

Dans les pages qui suivent, nous rendons compte de l'état d'avancement du projet. L'article est construit de la manière suivante. La section 2 décrit les objectifs généraux du projet et aborde la question de l'intérêt d'une annotation syntaxico-sémantique des contextes extraits de corpus spécialisés. La section 3 porte plus spécifiquement sur l'annotation manuelle réalisée dans deux domaines spécialisés, à savoir l'informatique et le réchauffement climatique. La section 4 porte sur la méthode d'automatisation de l'annotation envisagée et sur les étapes réalisées jusqu'à présent. Nous concluons en présentant quelques perspectives.

2 Pourquoi annoter les termes de nature prédicative ?

L'explication des propriétés syntaxico-sémantiques des termes dans les dictionnaires spécialisés et bases de données constituent une source de renseignements extrêmement utiles sur plusieurs plans. Nous nous intéressons plus spécifiquement aux termes de nature prédicative, et, dans une première étape, aux verbes.

Une description des propriétés syntaxico-sémantiques des verbes permet de mettre en évidence, en premier lieu, les constructions syntaxiques qu'ils admettent. Outre les propriétés des verbes eux-mêmes, la description s'intéresse au comportement de leurs participants, à savoir les actants (également appelés *arguments*) et les circonstants: [2] leur nombre, leur rôle sémantique, les modalités de leur combinatoire avec le verbe, etc.

2.1 Annotation syntaxico-sémantique des structures actancielles de termes

Comme nous l'avons signalé dans l'introduction, l'enrichissement des bases de données terminologiques envisagé ici prend la forme d'annotations insérées formellement dans les contextes dans lesquels apparaissent les termes. La méthodologie d'annotation s'inspire fortement de celle développée dans le cadre du projet FrameNet (Ruppenhofer 2002; FrameNet 2008),[3] mais s'en distingue sur certains plans. Une partie de ces distinctions seront évoquées plus loin dans l'article.

La figure 1 montre deux contextes dans lesquels les propriétés de termes sont explicitées : le premier illustre le terme d'informatique *affecter* et ses différents participants; le deuxième est extrait d'un corpus du changement climatique.

Les éléments explicités sont les suivants :

1. Terme prédicatif faisant l'objet de l'annotation (dans la figure 1, *affecter*);
2. Participants et leur nature (actants ou circonstants) : les participants sont mis en évidence directement dans le contexte, puis énumérés dans un tableau récapitulatif (à la figure 1, les actants sont en caractères gras dans le contexte et apparaissent dans la partie supérieure du tableau récapitulatif; les circonstants apparaissent dans la partie inférieure du tableau);
3. Rôle sémantique du participant : le rôle sémantique (**Agent**, **Patient**, **Destination**, etc.) est indiqué dans le contexte et dans le tableau récapitulatif (la question des rôles sémantiques est abordée à la section 2.3);

[2] La distinction entre "actant" et "circonstant" s'appuie sur Mel'čuk (2004). Les actants sont définis comme des participants obligatoires et contribuent au sens d'unités lexicales de sens prédicatif. Les circonstants apparaissent dans des phrases et peuvent entretenir un lien syntaxique avec l'unité lexicale, mais ne contribuent pas au sens de l'unité en question.

[3] Signalons qu'un nombre croissant de travaux terminologiques ont recours à des représentations s'inspirant des Frames sémantiques et de la méthodologie utilisée pour élaborer FrameNet (2008). Nous pouvons citer Dolbey *et al.* (2006) Faber *et al.* (2006) et Schmidt (à paraître).

4. Fonction syntaxique du participant : la fonction syntaxique (sujet, objet, complément, modificateur) du participant est indiquée dans le tableau récapitulatif;
5. Groupe syntaxique du participant : le groupe syntaxique (SN, SP, SAdv, Prop) du participant est également donné dans le tableau récapitulatif.

Les *développeurs* peuvent AFFECTER **une *valeur* à cette *variable***.

AFFECTER 1		
Actants		
Agent	Sujet (SN) (1)	développeur
Patient	Objet (SN) (1)	valeur
Récipient	Complément (SP-à) (1)	variable

L'*élévation* du niveau de la mer AFFECTERA **les *écosystèmes* des mangroves** en éliminant leurs habitats actuels et en créant des zones inondées par les marées vers lesquelles certaines espèces de mangrove pourraient se déplacer.

AFFECTER 1		
Actants		
Cause	Sujet (SN) (1)	élévation
Patient	Objet (SN) (1)	écosystème
Autres		
Mode	Complément (Prop) (2)	en éliminant leurs habitats actuels en créant des zones inondées par les marées vers lesquelles certaines espèces de mangrove pourraient se déplacer

Figure 1. Contextes annotés contenant le verbe *affecter*

2.2 Rôles sémantiques

Les annotations sont articulées autour des rôles sémantiques des participants (actants et circonstants) d'un terme prédicatif. Le rôle sémantique est défini comme le lien partagé par le participant et le terme prédicatif (**Agent**, **Patient**, **Instrument**, **Récipient**). Malgré le caractère subjectif qu'attribue parfois la littérature à la notion de « rôle sémantique », la représentation des participants par des étiquettes de rôles se révèle un outil efficace pour rendre compte des participants partageant le même lien avec des unités prédicatives différentes (Fillmore 1968; FrameNet 2008; VerbNet 2008).

L'exemple (1) montre de quelle manière les actants de termes de l'informatique sont représentés au moyen d'étiquettes de rôles. Bien que la position syntaxique des actants puisse varier, leur rôle restera le même.

(1) COMPILER$_{1a}$: **Instrument** ~ **Patient** (ex. *un compilateur compile du code*)
COMPILER$_{1b}$: **Agent** ~ **Patient** avec **Instrument** (ex. *un programmeur compile du code au moyen du compilateur x*)
COMPILABLE$_1$: **Patient** ~ (ex. *du code compilable*)
COMPILATEUR$_1$: ~ utilisé par **Agent** pour intervenir sur **Patient** (ex. *compilateur de Java*)

L'exemple (2) montre que l'annotation en rôles permet de mettre en évidence des liens de synonymie, de quasi-synonymie et d'opposition entre termes.

(2) STOCKER$_{1a}$: **Destination** ~ **Patient** (ex. *des végétaux stockent du CO_2*)
PIÉGER$_1$: **Destination** ~ **Patient** (ex. *le permafrost piège du dioxyde de carbone*)
EMPRISONNER$_1$: **Destination** ~ **Patient** (ex. *l'atmosphère emprisonne plus de chaleur*)

LIBÉRER$_{1a}$: **Source** ~ **Patient** (ex. *les eaux chaudes libèrent de la chaleur*)

Les étiquettes de rôles sémantiques auxquelles nous avons recours s'apparentent aux étiquettes utilisées pour représenter les éléments d'un Frame (FE) dans FrameNet (2008). Toutefois, elles s'en distinguent en ce sens que nous tentons de définir un nombre limité d'étiquettes qui s'appliqueront à l'ensemble des termes dans un domaine spécialisé (et non uniquement à l'intérieur d'un seul Frame).[4]

2.3 Schéma d'annotation XML

Le processus décrit en 2.1 est effectué en ajoutant des balises XML aux exemples d'utilisation et de descriptions tirés de corpus spécialisés. Afin de systématiser les différents types de balises selon les éléments décrits en 2.1 ainsi que leur imbrication, nous avons défini un schéma XML pour l'ensemble des unités lexicales que nous avons annotées. La figure 2 donne l'extrait du Schéma correspondant à l'annotation d'un participant. Chaque participant est identifié par son type (Actant ou circonstant) et son rôle (**Agent**, **Patient**, etc.) et contient une description de la fonction syntaxique (Objet, Sujet, etc.) exprimée en terme d'un groupe syntaxique (SN, SP, etc.) qui décrit la réalisation de cet actant par du texte comprenant une réalisation syntaxique (ceci est indiqué par *mixed*). À chaque réalisation est associée un identificateur *ref* qui permet de relier plusieurs occurrences (ex. une référence pronominale ou anaphorique) d'un même actant.

```
element-participant = element participant {
   attribute type {TypeParticipant},
   attribute role {RoleParticipant},
   element-fonction-syntaxique
}
element-fonction-syntaxique = element fonction-syntaxique {
   attribute nom {NomFonctionSyntaxique},
   attribute cas {CasFonctionSyntaxique}?,
   element-groupe-syntaxique
}
element-groupe-syntaxique = element groupe-syntaxique {
   attribute nom {NomGroupeSyntaxique},
   attribute preposition {text}?,
   attribute particule {text}?,
   mixed {element-realisation}
}
element-realisation = element realisation{
   attribute lemme {text}?,
   attribute etiquette {text}?,
   attribute ref {xsd:IDREF}?,
   attribute reflex {xsd:IDREF}?,
   text
}
```

Figure 2: Schéma XML (RelaxNG forme compact) des annotations d'un participant

Selon ce schéma, les participants dans le contexte **`Vous pouvez aussi, à ce moment-là, abandonner l'installation`**, sont annotés en XML comme indiqués à la figure 3. Cette annotation

[4] Ce faisant, nous nous opposons assez radicalement aux hypothèses formulées dans le cadre du modèle des Frames Sémantiques. Nous croyons effectivement qu'il est possible de rendre compte des structures actancielles de termes d'un domaine au moyen d'un nombre plus limité d'étiquettes s'apparentant davantage aux étiquettes proposées par Fillmore (1968). En outre, les unités lexicales ne sont pas regroupées dans des Frames et, par conséquent, ne sont pas hiérarchisées entre elles. Les hiérarchies proposées par FrameNet permettent de rendre compte de parentés sémantiques existant entre unités lexicales, mais appartenant à des Frames distincts.

permet une identification fine des participants: **`Vous`** comme **Agent**, **`à ce moment-là`** comme *Circonstant* avec rôle **Temps** et **`l'installation`** comme **Patient**. Le contenu de la balise `participant` donne des informations syntaxiques sur les actants. En utilisant un éditeur spécialisé pour XML, l'annotation peut être effectuée rapidement en sélectionnant les participants avec la souris et en choisissant les balises appropriées dans des menus. La définition d'un schéma autorise l'annotateur à n'afficher dans les menus que les balises qui, à chaque étape, font en sorte que l'annotation respecte le schéma.

L'utilisation de feuilles de styles appropriées offre la possibilité de générer plusieurs types de présentation des annotations : nous utilisons présentement des pages HTML qui distinguent les types d'actants par des codes de couleur (figure 1). Cette présentation permet également d'obtenir des statistiques sur le nombre d'occurrences des actants et leurs réalisations sur plusieurs contextes (figure 1). Ces informations permettent aux annotateurs de valider plus facilement leur travail car le code source XML est assez *touffu.* De plus, l'insertion de balises force une division des éléments de la phrase imposant à l'annotateur une lecture peu naturelle.

```
<participant type="Act" role="Agent">
  <fonction-syntaxique nom="Sujet">
    <groupe-syntaxique nom="SN">
       <realisation>Vous</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
 </participant> pouvez aussi,
 <participant type="Circ" role="Temps">
  <fonction-syntaxique nom="Complement">
    <groupe-syntaxique nom="SP" preposition="à">à ce
      <realisation>moment-là</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
 </participant>,
 <lexie-att>abandonner</lexie-att>
 <participant type="Act" role="Patient">
  <fonction-syntaxique nom="Objet">
    <groupe-syntaxique nom="SN">l'
      <realisation>installation</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
 </participant>
```

Figure 3 : Annotation XML (respectant les schémas de la figure 2) du contexte: `Vous pouvez aussi, à ce moment-là, abandonner l'installation.` La balise `lexie-att` non définie dans cet article, entoure l'occurrence de la lexie sous étude dans ce contexte.

3 Annotation des contextes dans deux domaines spécialisés

La méthodologie d'annotation manuelle des contextes comprend les étapes suivantes :

1. Constitution d'un corpus : la construction et la gestion du corpus suivent les recommandations établies au laboratoire du groupe ÉCLECTIK (Marshman 2003).
2. Extraction automatique de candidats termes simples : les candidats termes simples (nom, adjectif, verbe, adverbe) sont extraient par TermoStat (Drouin 2003). Ce logiciel d'acquisition automatique de termes s'appuie sur une approche contrastive, c'est-à-dire qu'il exploite une méthode de mise en opposition de corpus spécialisés et non spécialisés. Dans le présent travail, nous avons opposé un corpus de référence composé de textes journalistiques (Le Monde, année 2002) à un corpus d'informatique, d'une part, et à un corpus de changement climatique, d'autre part.

3. Analyse de la liste de candidats-termes : les listes de candidats générées par l'extracteur sont ensuite étudiées par des terminologiques qui ne retiennent que les unités étant de nature terminologique. Cette analyse nécessite l'application de critères de validation de la part des terminologues. Puisque nous annotons actuellement des termes de nature verbale, nous avons choisi des verbes extraits par TermoStat, puis validés par des analystes.
4. Choix de contextes dans lesquels apparaissent les termes : pour chaque terme sélectionné, nous prélevons entre 15 et 20 contextes. Pour préserver la représentativité du corpus, nous évitons de prendre trop de contextes figurant dans un même texte. De plus, à moins que les phrases ne soient vraiment trop longues, nous les prenons au complet afin d'inclure le maximum de participants. Il arrive que dans certaines phrases, un participant soit exprimé sous une forme anaphorique. Dans ces cas-là, nous prélevons également la phrase qui fait référence à ce participant (généralement la phrase qui précède). De cette façon, au moment de l'annotation, il est possible de relier l'anaphore au mot qu'elle renvoie.
5. Annotation dans le Schéma XML décrit à la section 2.3 : l'annotation est réalisée en deux étapes (première annotation et révision par une personne différente).

Actuellement, nous annotons des contextes tirés de deux corpus spécialisés distincts, à savoir un corpus d'informatique et un corpus portant sur le changement climatique. Les deux domaines abordant des thématiques assez différentes (concepts techniques dans le corpus d'informatique et concepts scientifiques parfois abordés sous un angle social dans le corpus sur le changement climatique), il apparaît intéressant de mettre au point le modèle d'annotation en les comparant l'un à l'autre.

3.1 Annotation des contextes : informatique

Les premières annotations manuelles ont été réalisées pour des verbes appartenant au domaine de l'informatique (ex. *abandonner, cliquer, configurer, joindre, télécharger*). Les verbes étaient déjà répertoriés dans une base de données terminologiques décrivant leur structure actancielle (la base de données porte le nom de DiCoInfo, Dictionnaire fondamental de l'informatique et de l'Internet[5]). Les annotateurs pouvaient donc se reporter à ces descriptions afin de distinguer les actants des circonstants dans les contextes réels.

Les contextes (rappelons que de 15 à 20 contextes sont retenus pour chaque unité lexicale annotée) sont extraits d'un corpus de textes rédigés en français d'environ 2 millions de mots. Les niveaux de spécialisation des textes varient, mais de manière générale, ces textes sont de nature didactique ou de vulgarisation et sont écrits, pour la plupart, par des spécialistes.

Les annotateurs disposaient d'une liste de rôles sémantiques préalablement définis dans la base de données pour étiqueter les actants (ex. **Agent**, **Patient**, **Instrument**, **Destination**, **Source**, **Assaillant**). Toutefois, puisque l'annotation des contextes prend également en compte les circonstants, nous avons dû définir de nouveaux rôles dont nous illustrons l'application ci-dessous :

But : *Pour lancer cette commande*
Environnement : *Sous Windows*
Temps : *à ce moment-là*
Voie : *via le bus de données*

Jusqu'à présent, nous avons annoté entre 15 et 20 contextes pour approximativement 122 unités lexicales. Ces contextes ont exigé la définition d'un total 32 étiquettes de rôles sémantiques.

3.2 Annotation des contextes : changement climatique

Pour les contextes du changement climatique, contrairement aux contextes de l'informatique, l'annotation précède la création des fiches, et ce principalement pour deux raisons. Premièrement, l'élaboration du dictionnaire de l'informatique est antérieure au projet d'annotation. Deuxièmement, débuter par l'annotation semble une démarche progressive et logique. En ce sens qu'elle permet aux terminologies de

[5] Le DiCoInfo est accessible à l'adresse suivante : http://olst.ling.umontreal.ca/dicoinfo/.

définir la structure actancielle des termes en se basant sur des données contextuelles. Les fiches sont par conséquent composées d'après les faits de langues observés en corpus – les contextes ne sont pas choisis en fonction des fiches.

Le corpus sur lequel nous travaillons compte approximativement 1 020 000 mots. Pour la sélection des textes, nous nous sommes appuyés principalement sur cinq critères : 1) les auteurs des textes sources sont des spécialistes ou du moins des personnes connaissant très bien le domaine; 2) les textes proviennent de sites Internet ou de publications reconnus dans le domaine; 3) les textes sont diversifiés au niveau de la spécialisation (technique, scientifique, vulgarisé, etc.); 4) les textes proviennent de différents types de documents (manuels, périodiques, sites Internet, etc.); 5) l'équilibre entre la taille et le nombre de documents est respecté.

Parmi les verbes que nous annotons, certains ont la même forme que les unités lexicales utilisées en informatique, mais la plupart ont des acceptions distinctes : *affecter*, *analyser*, *calculer*, *emmagasiner*, *générer*, *modifier*. Ce recoupement, permettra de vérifier dans quelle mesure le domaine influence les propriétés linguistiques de ces formes.

Dès à présent, il nous est possible d'identifier des rôles sémantiques qui se manifestent avec plus de régularité dans le domaine du changement climatique que dans celui de l'informatique. Au nombre des rôles associés aux actants, **Cause** et **Destination** sont particulièrement fréquents (figures 1 et 2.2). Par ailleurs, la liste des rôles associés aux circonstants à, pour le moment, été augmentée par des rôles à caractère spatial, quantitatif et temporel :

Direction : *vers le nord*
Coût : *à un coût se situant entre 50 et 180 dollars É.-U. par véhicule*
Valeur : *de 50 à 70 %*
Durée : *durant une courte période hivernale*

4 Annotation automatique des contextes : premières étapes

La tâche d'annotation manuelle décrite à la section 3 est très fastidieuse : l'annotation des contextes d'une unité lexicale exige 2 heures en moyenne. Notre corpus pour l'instant est constitué de 105 unités lexicales et nous avons annoté manuellement environ 3321 contextes.[6] L'annotation de ces 3321 contextes a nécessité environ 4 mois de travail. Pour accélérer l'annotation, nous proposons une méthode automatique d'apprentissage des traits des actants en nous basant sur un système de classification entraîné sur notre corpus de données annotées manuellement avec les rôles sémantiques tels qu'**Agent**, **Patient**, **Destination**, **Instrument**, etc.

La tâche consiste à annoter les participants des unités lexicales en rôles sémantiques. Ces participants sont de deux types : actants et circonstants. L'annotation en rôles sémantiques portera d'abord sur les actants, car la distinction entre les circonstants est plus complexe. Dans notre travail, trois tâches sont considérées : 1) identification des participants; 2) distinction entre les actants et les circonstants; 3) assignation de rôles sémantiques aux actants. Nous utilisons les liens syntaxiques d'unité lexicale avec les autres mots de la phrase comme traits caractéristiques ou « features » pour réaliser ces tâches principales. Les liens syntaxiques entre les unités de la phrase sont identifiés à l'aide de l'analyseur syntaxique *Syntex* (Bourigault *et al.* 2005).

On interprète la combinaison de l'analyse produite par *Syntex* et les liens entre l'unité lexicale et ses participants identifiés lors de l'analyse manuelle au moyen d'un graphe (voir figure 4).

[6] La méthode d'annotation a été mise au point à partir des contextes relevés pour des verbes d'informatique et non ceux du corpus de réchauffement climatique. Le nombre total d'unités lexicales retenues est moins élevé que le nombre total d'unités annotées puisque le travail d'annotation manuelle s'est poursuivi en parallèle.

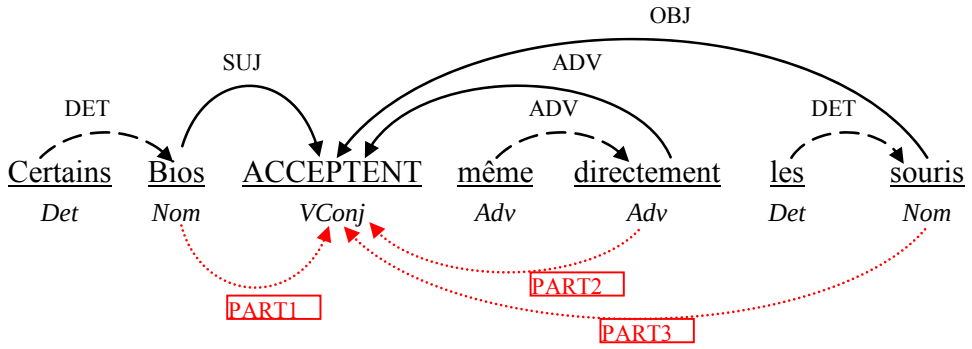


Figure 4 : Graphe de combinaison et règles déduites. Les flèches au-dessus des mots correspondent aux liens donnés par Syntex. Nous indiquons en pointillé les flèches qui ne seront pas considérées. Les flèches sous les mots correspondent aux participants annotés manuellement.

À partir d'un tel graphe, on extrait des règles d'identification de participants et leurs « features ». Ces règles affecteront le type Actant ou Circonstant à ces participants. Ces features sont aussi utilisés pour trouver les rôles sémantiques des participants identifiés. Sous le graphe, on indique une règle déduite par l'observation manuelle de plusieurs situations semblables.

Une règle a deux parties : la partie gauche, qui constitue les conditions de l'application, est composée des unités de la phrase avec les liens syntaxiques qui se trouvent au-dessus des unités et la partie droite qui représente les résultats de l'application où les participants sont indiqués sous les unités.

Dans une règle, une unité est décrite par ses traits. Un trait est écrit *<mot, Cat, Fonc>*, où *Cat* correspond à la catégorie grammaticale de l'unité et *Fonc* est sa fonction syntaxique. Dans une règle, une unité peut appartenir à plusieurs catégories grammaticales; dans ce cas, *Cat* est noté comme une alternative de plusieurs étiquettes syntaxiques, chacune d'elles séparées par un «|». Par exemple, le trait du mot de catégorie *Nom* ou *Pronom* est écrit *<mot, Nom|Pro, Role>*.

La figure 5 montre une partie des résultats obtenus lors de l'identification des participants au moyen de la stratégie décrite ci-dessus.

Lorsque la chaîne atteint la longueur maxi on recherche le premier espace en partant de la fin la ***partie*** gauche est imprimée puis ANNULÉE

Pour ABANDONNER le ***processus*** d installation à ce ***stade*** redémarrez l ordinateur et retirez la disquette d amorçage ou le CD ROM

ABANDONNEZ la ***copie*** ou la ***mise à jour***

Pour ce faire le ***système*** est capable de FERMER un ***environnement*** Classic devenu instable ou inopérant

Si l on veut protéger les ***informations*** ENREGISTRÉES sur une ***disquette*** de13 5 cm il suffit d obstruer cette encoche avec un morceau de papier adhésif

Dans les débuts de l informatique les ***programmeurs*** ÉCRIVAIENT des ***programmes*** ***en codant directement en langage machine***

Les ***disques*** à GRAVER sont conditionnés en plastique dur

La requête actuellement affichée a été modifiée et ***vous*** tentez de SORTIR sans la sauver

Certains ***BIOS*** ACCEPTENT même ***directement*** les ***souris***

Figure 5 Participants des unités lexicales identifiés automatiquement. Les unités lexicales sont en majuscules. Les mots en gras italique et soulignés sont leurs participants (Actants et Circonstants)

À partir d'un certain nombre d'exemples, nous avons dégagé une quarantaine de patrons d'identification de participants et nous les avons appliqués à l'ensemble du corpus (en dehors des

exemples que nous avions étudiés pour le développement de nos règles) soit 105 lexies sur 3321 contextes. En comparant les participants identifiés automatiquement et ceux qui avaient été identifiés manuellement, nous avons obtenu une précision de 75 % (nombre de participants pertinents retrouvés par rapport au nombre de participants total) et un rappel de 80 % (nombre de participants pertinents retrouvés par rapport au nombre de participants pertinents). Ces premiers résultats nous montrent que nous sommes sur la bonne voie et qu'il est possible d'accélérer le processus d'identification des actants au moyen de règles automatiques. Ces annotations pourraient servir de point de départ, quitte à ce qu'elles soient révisées par la suite.

Pour l'attribution des rôles sémantiques aux actants, certaines fonctions syntaxiques retournées par l'analyseur *Syntex* nous permettent d'affecter avec un bon niveau de confiance des rôles sémantiques aux actants de l'unité lexicale. Par exemple, les fonctions syntaxiques « Sujet » ou « Objet » nous permettent d'attribuer les rôles **Agent** ou **Patient** aux actants correspondants. Alors que pour les actants ayant ces fonctions, on arrive à leur affecter des rôles sémantiques, ce n'est pas toujours le cas pour d'autres fonctions. Par exemple, les actants prépositionnels occupant une fonction de Complément peuvent être identifiés à l'aide des liens syntaxiques mais l'attribution des rôles sémantiques ne pourra s'appuyer uniquement sur le type de fonction syntaxique renvoyé par Syntex. Dans ce cas, on utilise les mêmes features (dépendances syntaxiques et position du participant indiquant s'il apparaît avant ou après l'unité lexicale dans la phrase) de l'étape d'identification de participants auxquels on rajoute d'autres features tels le feature tête comme les prépositions *à*, *dans*, *sur*, *grâce à*, *avec* pour les prépositionnels ou le feature Lexie (unité lexicale) car le rôle d'un actant peut aussi dépendre de l'unité lexicale avec laquelle il est utilisé.

Dans une étape ultérieure, nous comptons utiliser des méthodes d'apprentissage machine pour nos traitements. Nous utiliserons des classificateurs automatiques qui s'appuieront sur les features cités ci-dessus. À terme, nous expérimenterons avec deux approches d'apprentissage : statistique (Gildea & Jurafsky 2002) et *instance-based learning* (Li *et al.* 2001).

5 Conclusion et perspectives

Dans cet article, nous avons voulu rendre compte d'un projet en cours visant à enrichir des bases de données terminologiques en y introduisant, plus spécifiquement, des renseignements de nature syntaxico-sémantique. Ce projet comporte un travail manuel important, mais nous espérons le réduire de manière considérable au moyen d'une méthode qui permettra de générer des annotations automatiquement, annotations qui seront par la suite révisées par des annotateurs humains.

Les résultats obtenus jusqu'à présent quant à l'identification automatique des participants sont prometteurs. Nous espérons pousser plus avant l'automatisation en procédant à l'attribution automatique de certains rôles, notamment ceux associés aux actants. Ce travail nécessite une définition rigoureuse des rôles décrits jusqu'à maintenant et une meilleure compréhension des structures syntaxiques dans lesquelles ils peuvent se retrouver. Naturellement, la démarche n'est pas linéaire, rien n'empêche une fois la rédaction des fiches commencée d'apporter des modifications aux contextes annotés. Autrement dit, le travail d'annotation automatique repose sur une analyse manuelle importante; de même, le traitement automatique des contextes – en repérant certaines erreurs – permet de systématiser des paramètres de l'annotation manuelle.

Remerciements

Les travaux de recherche présentés dans le présent article sont financés par le Conseil canadien en sciences humaines (CRSH) du Canada et par les Fonds québécois de recherche sur la société et la culture (FQRSC). Les auteurs aimeraient remercier Stéphanie Caron, Stéphanie Klebetsanis et Charlotte Tellier qui ont participé au travail d'annotation manuelle des contextes d'informatique.

Références

Binon, Jean, Serge Verlinde, Jan Van Dyck & Ann Bertels. 2000. *Dictionnaire d'apprentissage du français des affaires*. Paris: Didier.

Bourigault, Didier, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques & Sylvia Ozsdowska. 2005. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*. Dourdan, France.

DiCoInfo. Dictionnaire fondamental de l'informatique et de l'Internet. (http://olst.ling.umontreal.ca/dicoinfo/). Accessed 5 February 2009.

Dolbey, Andrew, Michael Ellsworth & Jan Scheffczyk. 2006. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In Bodenreider, Olivier (ed.). *Proceedings of KR-MED*, 87-94.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1): 99-117.

Faber, Pamela, Silvia Montero Martínez, María Rosa Castro Prieto, José Senso Ruiz, Juan Antonio Prieto Velasco, Pilar León Arauz, Carlos Márquez Linares & Miguel Vega Expósito. 2006. Process-oriented terminology management in the domain of Coastal Engineering. *Terminology*, 12(2): 189-213.

Fillmore, Charles J. 1968. The case for case. In Bach, Emmon & Robert T. Harns (eds.). *Universals in linguistic Theory*, New York: Holt, Rinehard & Winston, 1-88.

Fillmore, Charles J., Christopher R. Johnson & Miriam R.L. Petruck. 2003a. Background to FrameNet, In Fontenelle, Thierry (ed.). FrameNet and Frame Semantics. Special Issue of the *International Journal of Lexicography*, 16(3): 235-250.

Fillmore, Charles J., Miriam R.L. Petruck, Joseph Ruppenhofer & Abby Wright. 2003. FrameNet in Action: The case of attaching. *International Journal of Lexicography*, 16(3): 297-332.

FrameNet (http://framenet.icsi.berkeley.edu/). Accessed 6 February 2008.

Gildea, Daniel & Daniel Jurafsky. 2002. Automatic labelling of semantic roles. *Computational Linguistics*, 28(3): 245-288.

Lépinette, Brigitte & Françoise Olmo Cazevieille. 2007. Un dictionnaire syntaxico-sémantique bilingue de la zootechnie. *Cahier de linguistique*, 33(1): 83-102.

Li Jiamming, Lei Zhang & Yong Yu. 2001. Learning to Generate Semantic Annotation for Domain Specific Sentences. In *Workshop on Knowledge Markup and Semantic Annotation at the 1st International Conference on Knowledge Capture* (K-CAP 2001), October, Victoria, B.C., Canada

Marshman, Elizabeth. 2003. Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie. (http://www.olst.umontreal.ca/pdf/terminotique/corpusentermino.pdf). Accessed 8 February 2009.

Mel'čuk, Igor. 2004. Actants in semantics and syntax. I: actants in semantics. *Linguistics*, 42(1): 1-66.

Ruppenhofer, Joseph, Michael Ellsworth, Miriam R.L. Petruck, Christopher Johnson & Jan Scheffczyk. 2002. *FrameNet II: Extended Theory and Practice*. (http://framenet.icsi.berkeley.edu/). Accessed 6 February 2008.

Schmidt, Thomas. forthcoming. *The Kicktionary – A Multilingual Lexical Resources of Football Language.*

VerbNet (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html). Accessed 7 February 2008.

La mise en équivalence des adjectifs relationnels du domaine médical : étude du suffixe *–ionnel*.

François Maniez
Centre de Recherche en Terminologie et Traduction
Université Lumière-Lyon 2
86 rue Pasteur 69365 Lyon Cedex 7
francois.maniez@univ-lyon2.fr

Abstract

Nous nous proposons d'étudier l'utilisation d'un certain type d'adjectifs relationnels (ceux se terminant par *–ionnel*) dans le domaine de la médecine. La traduction de ces adjectifs en anglais pose divers problèmes pour les francophones, et notamment celui du choix entre une traduction quasi-littérale, utilisant un adjectif de suffixation équivalente *(-ion + -al*) où une complémentation prépositionnelle impliquant le nom dont l'adjectif est dérivé. Après avoir étudié la sélection de ces adjectifs dans deux ressources terminologiques bilingues, nous étudions l'emploi des dix adjectifs les plus fréquents et celui de leurs équivalents morphologiques anglais dans deux corpus comparables d'articles de recherche du domaine de la médecine, un corpus de langue française de huit millions de mots et la partie médicale du *Corpus of Contemporary American English* (accessible en ligne et totalisant 4,5 millions de mots). Nous tentons d'identifier les facteurs qui contribuent à l'utilisation du nom ou de l'adjectif dans les expressions anglaises candidates à la mise en équivalence des expressions françaises utilisant ces adjectifs.

1 Introduction

Dans un ouvrage publié il y a plus d'une trentaine d'années, le professeur Sournia (1974) vitupérait un certain nombre d'utilisations de l'adjectif dont la traduction de locutions d'origine anglo-saxonne était responsable, citant entre autres « traitements institutionnels » ou « profils comportementaux ». Il relevait également l'usage abusif de l'adjectif par certains de ses confrères, regrettant l'ambiguïté de formulations telles que « chirurgien infantile » ou « chirurgien cardiaque », hypallages produites par l'utilisation de la totalité du nom d'une spécialité pour en désigner le spécialiste (l'anglais produit ainsi *criminal lawyer* à partir de *criminal law*, le droit pénal).

Si ces usages ne soulèvent plus guère d'objections aujourd'hui, le combat contre l'ambiguïté que peut générer le calque systématique de la dérivation adjectivale de l'anglais reste plus que jamais justifié. L'adjectif « comportemental » est de fait entré dans la langue française il y a maintenant soixante ans, mais les pourfendeurs d'anglicismes dans la prose scientifique ont l'occasion de se battre actuellement contre d'autres emprunts, tels « développemental » ou « volitionnel », dont l'usage n'est pas encore consacré par les dictionnaires de la langue générale, mais largement attesté par les moteurs de recherche (le premier étant utilisé depuis une vingtaine d'années et comptant actuellement plusieurs dizaines de milliers d'occurrences sur la Toile).

Nous avons mentionné le cas de l'adjectif « comportemental », dans lequel un néologisme français résulte de l'adoption d'une dérivation adjectivale de l'équivalent anglais (behavioral). Cependant, la forma-

tion de nouveaux adjectifs français pour traduire des expressions équivalentes venant de l'anglais est parfois justifiée par un phénomène syntaxique typique des langues germaniques, la prémodification nominale par un nom adjectival. La traduction par le complément de nom est souvent utilisée pour ce type de structure, mais l'utilisation de l'adjectif offre plusieurs avantages : d'une part, elle confère à l'expression utilisée une consonance plus technique par son aspect figé ; d'autre part, elle évite un choix parfois épineux, celui entre article défini ou indéfini pour introduire le complément de nom. Ainsi, *fracture reduction* sera traduit par « réduction fracturaire », *fracture site* par « foyer fracturaire », alors que l'adjectif « fracturaire » n'a pas d'équivalent dans le lexique anglais. On peut présumer que de tels termes donnent progressivement naissance à des expressions pour lesquelles l'usage de l'adjectif ne paraît pas indispensable. Ainsi, « risque fracturaire » supplante peu à peu « risque de fracture » comme traduction de *fracture risk*. Il faut toutefois signaler que l'utilisation de l'adjectif, fustigée par certains comme jargonneuse, permet de gagner en concision dans les cas où la prémodification nominale de l'anglais exprime des relations plus complexes que celles généralement dévolues au complément de nom : c'est le cas de collocations comme « lésion fracturaire » ou « fragment fracturaire », dans lesquelles l'adjectif signifie « résultant d'une fracture, consécutif à une fracture ».[2]

La structure du nom adjectival pose donc au traducteur et au terminographe dont le français est la langue cible le problème du choix entre deux structures possibles : celle du complément de nom et celle qui utilisera l'adjectif formé par dérivation à partir du nom français correspondant, si cet adjectif existe. Ce choix est déterminé par l'usage : *cell repository* est traduit par « banque de cellules », *cell wall* par « paroi cellulaire », cet usage étant figé dans le cas de ces deux collocations. Quand un choix existe, cet usage dépend toutefois du locuteur et du contexte de communication. Un médecin écrira (et dira dans certains contextes) « cancer mammaire » ou « infarctus myocardique » là où le non-spécialiste parlera de cancer du sein et d'infarctus du myocarde. Il s'agit là de deux exemples pour lesquels l'anglais ne présente pas réellement de variation, puisque la première expression utilise quasi-exclusivement une prémodification nominale (*breast cancer*) et l'autre un adjectif (*myocardial infarction*). Toutefois, on peut supposer que l'adjectif relationnel anglais joue lui aussi ce rôle de démarcation entre le discours du spécialiste et celui du non-spécialiste.[3]

Nous examinerons ici l'emploi de certains adjectifs relationnels dans un corpus d'articles de recherche du domaine de la médecine totalisant huit millions de mots et tenterons de déterminer quels sont les facteurs qui contribuent à l'utilisation du nom ou de l'adjectif dans les expressions anglaises qui sont des équivalents de traduction des expressions françaises utilisant ces adjectifs. Nous avons choisi d'étudier ceux se terminant par –ionnel car leur traduction littérale en anglais n'est pas toujours possible, la suffixation en –*el* à partir d'un nom déverbal semblant s'appliquer moins fréquemment. Par ailleurs, il semblerait que l'utilisation de ces adjectifs en anglais soit essentiellement le fait de locuteurs non natifs.

2 Propriétés syntaxiques des adjectifs relationnels

On sait que la majorité des termes employés en anglais de spécialité sont de longueur 2 (Frantzi et al 1999), et sont généralement formés selon un modèle comprenant un élément modificateur (adjectif ou nom) qui précède le nom constituant le « nœud » du terme. Le passage d'une langue utilisant la prémodification (comme les langues germaniques) à une langue utilisant la post-modification (langues romanes) par le biais d'un groupe prépositionnel nécessite en effet l'explicitation de cette relation. Toutefois, les adjectifs dits « relationnels »[4] peuvent également traduire cette prémodification. Les

[2] Notons au passage que *fracturaire* et ses divers composés (*anti-fracturaire, inter-fracturaire, micro-fracturaire, péri-fracturaire, post-fracturaire, pré-fracturaire*) totalisent une dizaine de milliers d'emplois sur la Toile, mais que l'adjectif *fracturaire* est absent du Grand Robert.

[3] A titre d'exemple, on dénombre dans le corpus CCAE (décrit plus bas) 14 occurrences de *kidney function* et 98 occurrences de *renal function* dans la partie consacrée aux articles de médecines, alors que les totaux sont respectivement de 36 et 3 pour la partie consacrée aux magazines non spécialisés.

[4] Certains auteurs les nomment également adjectifs argumentaux. En anglais, ils sont alternativement décrits par les appellations *relative adjectives* et *relational adjectives*, et certains (Warren 1984) parlent de *classifying adjectives*.

relations entre les deux noms concernés par la prémodification peuvent être de natures multiples. Le prémodificateur peut signifier la localisation anatomique (*back pain* → douleur dorsale ou dorsalgie, *brain stem* → tronc cérébral), la fonction (*sweat glands* → glandes sudoripares), la cause (*heat rash* → erythème calorique) ou la forme (*sickle cell* → cellule falciforme).[5] Ce peut également être un nom désignant une substance présente dans l'organisme. (*serum iron* → fer sérique, *plasma half-life* → demie-vie plasmatique).

Mélis-Puchulu (1991) considère que seul le critère morphologique discrimine de façon certaine les adjectifs relationnels, qui sont des adjectifs dénominaux dérivés d'un nom commun (présidentiel) ou d'un nom propre (suédois, hugolien). Les suffixes donnant naissance à ces adjectifs sont nombreux : Mélis-Puchulu cite ainsi -acé, -ain, -aire, -al, -ard, -é, -el, -esque, -eux, -ien, -ier, -in, -iaque, -ique, -iste, -ite, -ois, -ol, -ote et -u (mais -if et -oire ne sont pas mentionnés). En anglais, la liste est moins longue, les principaux suffixes de l'anglais médical étant -al, -ar, -ate, -ary, -iac, -ic, -ive, -ory, et -ous. Fradin (2008) fait toutefois remarquer que parmi les adjectifs dénominaux figurent certains adjectifs qui ne sont pas (ou plus) relationnels comme « occasionnel » ou « personnel ».

Une des propriétés syntaxiques des adjectifs relationnels français n'est pas transférable à l'anglais, celle de l'impossibilité de l'antéposition par rapport au nom (*les présidentielles élections). Les deux autres propriétés le plus souvent évoquées sont l'absence d'emplois attributifs (**The elections are presidential*) et l'absence de gradation (**It is very presidential*). Fradin souligne également que certains adjectifs peuvent être considérés comme relationnels ou non en fonction de leurs usages, ce qui explique pourquoi les emplois attributifs sont possibles dans certains cas (Son tempérament est lymphatique) et non dans d'autres (*Ce canal est lymphatique).

On sait depuis L'Homme (2004) que les adjectifs relationnels constituent la majorité des adjectifs de la terminologie médicale. Certains auteurs (Daille 2001) se sont penchés sur le problème de leur identification automatique en corpus. En anglais, la relativement faible représentation de certains noms par rapport aux adjectifs qui en sont dérivés (*clinic/clinical*) peut rendre cette identification difficile, ainsi que l'absence de relation morphologique entre nom et adjectif (*dental/tooth, cardiac/heart*).[6] Par ailleurs, certains adjectifs ont plusieurs acceptions si bien qu'il est possible qu'ils soient employés attributivement en corpus tout en ayant d'autres emplois relationnels (*Among other aspects of the plan, the investigational staff was significantly increased. / Two of the current NTL ICD systems described above are still investigational*). Les emplois attributifs sont même relativement fréquents pour désigner certains types de patients (*Guntheroth then showed that […] 40% of patients are bacteremic just after dental extractions. / This entity was noted in siblings who were hypertensive and who also had a marked concordance for dyslipidemia*). Le seul critère fiable d'identification des adjectifs relationnels semble celui de l'ordre des adjectifs, les adjectifs relationnels n'étant pas séparés du nom qu'ils modifient par un adjectif qualificatif (*This is not to say that one should not perform a complete neurological examination, rather that one should make sure that the basics are covered first. / *This is not to say that one should not perform a neurological complete examination […]*).

---

[5] La tendance à la transposition du nom adjectival provient en partie du fait que certains adjectifs français n'ont pas de dérivation équivalente en anglais. L'examen des suffixes de dérivation les plus productifs en français (-aire, -al, -eux, -ien et surtout -ique) révèle la présence de plus de 400 adjectifs différents dans un corpus d'articles de médecine de 750 000 mots Bon nombre de ces adjectifs ne peuvent se traduire en anglais qu'à l'aide de la prémodification par un nom adjectival. Pour ne considérer que le cas des dérivations par le suffixe -aire, des adjectifs tels que « canalaire », « fracturaire », « métaphysaire », « ovalaire », « plasmocytaire », « pubertaire », « séquellaire » ou « tissulaire » n'ont pas d'équivalent anglais. Cette absence est encore plus frappante dans le domaine des adjectifs composés, où l'anglais est beaucoup moins productif que le français : les équivalents anglais font défaut pour des adjectifs tels que « cholangiocellulaire », « dorso-lombaire », « endo-biliaire », « fémoro-patellaire » ou « fibro-glandulaire ».

[6] Le lien morphologique peut être davantage obscurci quand c'est l'ensemble d'un groupe nominal qui donne naissance à un adjectif. Ainsi, on relève des formes telles que *coronary arteri<u>al</u> vasospasm* (= vasospasm of the coronary arteries), *jugular ven<u>ous</u> pressure* (= pressure of the jugular veins) et même *inferior vena cav<u>al</u> obstruction* (obstruction of the inferior vena cava).

3 Sélection des adjectifs dénominaux dans les dictionnaires terminologiques

Les expressions nominales contenant des adjectifs relationnels sont faiblement répertoriées dans les dictionnaires spécialisés, probablement du fait qu'elles sont perçues comme appartenant à la phraséologie du domaine plutôt qu'à sa terminologie. Pour reprendre deux des exemples cités plus haut, on ne relève que deux expressions contenant l'adjectif « lésionnel » dans le Grand Dictionnaire Terminologique (syndrome lésionnel, topographie lésionnelle) et une mention de l'adjectif « incisionnel », le terme « hernie incisionnelle » étant mentionné comme synonyme d'éventration postopératoire (*incisional hernia*).

Adjectif	Partie anglaise	Partie française
nutrition(al/nel)	46	45
position(al/nel)	32	3
junction(al/nel)	3	3
transcription(al/nel)	2	2
gestation(al/nel)	1	2
confusion(al/nel)	1	4
lesion(al/nel)	0	2
transfusion(al/nel)	0	2
tension(al/nel)	0	2
ejection (al/nel)	0	1
retention(al/nel)	0	1

Tableau 1. Nombre de termes bilexicaux formés à partir d'adjectifs de la terminologie médicale en *–ion(al* dans le Grand Dictionnaire Terminologique

Nous avons également consulté une autre ressource terminologique médicale bilingue, la version bilingue du *Dorland's Pocket Medical Dictionary*. Comme dans le Grand Dictionnaire Terminologique, très peu d'adjectifs font l'objet d'une entrée distincte (*accessional*, *functional*, *locoregional*, *professional*, *progestational*, *provisional*, *successional*), la majorité des entrées contenant ces adjectifs étant de forme ADJ-N (*constitutional disease*, *emotional deprivation*, *functional murmur*). Les traductions de l'adjectif dans ces expressions sont majoritairement littérales, *-ional* donnant *-ionnel*, avec une modification vocalique éventuelle (*functional*/*fonctionnel*). Les adjectifs présents dans les groupes nominaux formant les vedettes de ces entrées sont les suivants : *appositional*, *constitutional*, *positional*, *(dys/non)functional*, *gestational*, *ideational*, *incisional*, *junctional*, *nutritional*, *oppositional*, *rational*, *transactional*, *volitional*.

Les cas d'absence de littéralité de la traduction font apparaître plusieurs phénomènes. Morphologiquement, on peut distinguer les cas dans lesquels un autre suffixe suffixe adjectival est utilisé (*accessional teeth* : dents accessoires, *fractional distillation* : distillation fractionnée, *progestational* : progestatif, *recreational drug* : drogue récréative) de ceux dans lesquels l'absence de dérivation à partir d'un cognat français provoque l'adoption d'une autre base nominale (*accessional* : acquisitionnel, *inhalational anthrax* : charbon pulmonaire, *licensed vocational nurse* : infirmier professionnel diplômé, *organic delusional syndrome* : trouble délirant organique, *postconcussional disorder* : syndrome commotionnel).

On observe également un bon nombre de transpositions[7] de l'adjectif vers un groupe prépositionnel (*depletional hyponatremia* : hyponatrémie par déplétion, *dilutional hyponatremia* : hyponatrémie par dilution, *excisional biopsy* : biopsie d'excision, *exertional headache* : céphalée d'effort, *positional vertigo* : vertige de position, *torsional diplopia* : diplopie de torsion, *transitional epithelium* : épithélium de transition).[8]

[7] On entend par transposition un changement de catégorie grammaticale lors du processus de traduction.

[8] L'utilisation du nom ne provient pas toujours d'une absence de l'adjectif du lexique. Ainsi, *transitionnel* existe depuis le XIX° siècle, et *dilutionnel* ou *excisionnel* sont d'usage fréquent sur la Toile, *biopsie excisionnelle* étant même plus fréquemment utilisé que ses équivalents à base de complément prépositionnel. C'est également le cas pour *vertige positionnel* et *épithélium transitionnel*.

Les cas de variation sont rares, et concernent souvent une distinction entre des emplois spécialisés et non-spécialisés. Ainsi, *successional* est traduit par « successif », et *successional teeth* par « dents de remplacement »; *provisional* est traduit par « provisionnel » alors que *provisional denture* donne « dentier provisoire ». L'adjectif *occupational*, très employé en terminologie (plus de 100 entrées dans le Grand Dictionnaire Terminologique) donne lieu à des traductions adjectivales (*occupational disease* : maladie professionnelle), par transposition (*occupational medicine* : médecine du travail) ou par l'utilisation du formant grec ergo- (*occupational therapy* : ergothérapie).

4 Corpus utilisés

Pour étudier les emplois des adjectifs relationnels français et de leurs possibles équivalents anglais, nous avons eu recours à deux corpus unilingues comparables. L'un est constitué d'articles de recherche dans le domaine médical tirés des revues disponibles sur la base de données Science Direct, et les droits de reproduction et d'utilisation dans le cadre d'un projet de recherche ont été accordés par les éditions Elsevier pour la quasi-totalité des publications. Le corpus a été étiqueté en partie du discours par l'analyseur Cordial, et compte actuellement huit millions de mots.[9] Nous ferons référence à ce corpus sous le nom de MED_FR dans les lignes qui suivent.

Le corpus anglais est la partie médicale du *Corpus of Contemporary American English* (accessible à l'adresse http://www.americancorpus.org/, et désigné ci-dessous par le sigle CCAE). Le corpus, qui totalise 385 millions de mots, contient une partie composée exclusivement d'articles de recherche universitaire (nommée *Academic*), elle-même divisée en plusieurs sous-sections dont une partie médicale qui totalise un peu plus de 4,5 millions de mots (les cinq parties principales du corpus, intitulées *Spoken*, *Fiction*, *Magazine*, *Newspaper* et *Academic* contiennent chacune entre 72 et 80 millions de mots).

L'étude de la fréquence comparée des divers suffixes adjectivaux de l'anglais dans la partie *Academic* et sur l'ensemble du corpus a révélé une forte présence du suffixe adjectival t -*ional* (42,4%) dans la partie *Academic* du corpus. Nous avons choisi d'étudier la présence de celui-ci en raisons des problèmes de traduction mentionnés plus haut pour les adjectifs dénominaux.

5 Les adjectifs en –*ionnel* du corpus MED_FR

Un examen rapide des groupes nominaux de patron syntaxique Nom-Adjectif formés à partir d'adjectifs en –*ionnel* les plus fréquemment employés dans MED_FR révèle la forte représentation d'un petit nombre d'adjectifs (*transfusionnel*, *fonctionnel*, *professionnel*, *nutritionnel*) :

Fréquence	Nom	Adjectif
293	sécurité	transfusionnelle
251	résultats	fonctionnels
194	asthme	professionnel
164	incidents	transfusionnels
143	résultat	fonctionnel
136	activité	professionnelle
133	traitement	conventionnel
117	explorations	fonctionnelles
115	risque	transfusionnel
115	état	nutritionnel
111	dossier	transfusionnel
110	accord	professionnel

[9] Les diverses spécialités médicales sont à peu près également représentées, comme en témoigne la liste des revues utilisées : Annales de Cardiologie et d'Angéiologie, Annales de Chirurgie, Annales de Chirurgie Plastique Esthétique, Annales Françaises d'Anesthésie et de Réanimation, Annales Médico-psychologiques, Annales de Réadaptation et de Médecine Physique, Annales d'Urologie, Médecine et Maladies Infectieuses, Néphrologie & Thérapeutique, Revue Française d'Allergologie et d'Immunologie Clinique, Revue de Médecine Interne, Revue du Rhumatisme, Transfusion Clinique et Biologique.

99	arbre	décisionnel
96	chirurgie	conventionnelle
96	pratiques	professionnelles
87	acte	transfusionnel
78	bilan	lésionnel
76	exploration	fonctionnelle
71	support	nutritionnel
69	incident	transfusionnel
68	plan	fonctionnel
63	séquelles	fonctionnelles
61	besoins	transfusionnels
58	chaîne	transfusionnelle
56	âge	gestationnel
55	gêne	fonctionnelle
54	troubles	mictionnels
54	asthmes	professionnels
53	seuil	transfusionnel
51	récupération	fonctionnelle
50	pronostic	fonctionnel

Tableau 2. Groupes nominaux de patron syntaxique Nom-Adjectif formés à partir d'adjectifs en -ionnel les plus fréquemment employés dans MED_FR

La comparaison de ces emplois adjectivaux avec l'emploi de noms ou d'adjectifs dans les structures anglaises pouvant servir d'équivalent de traduction doit-elle être opérée sur la seule base de la fréquence ? Comme souvent en pareil cas, les cas les plus statistiquement représentatifs ne sont pas les plus intéressants à étudier. Ainsi, bon nombre d'entre eux (qui ne sont parfois pas strictement relationnels) se traduisent systématiquement en anglais par l'adjectif morphologiquement correspondant (fonctionnel/*functional*, professionnel/*professional*, conventionnel/*conventional*). D'autres n'ont pas d'adjectif morphologiquement correspondant (c'est le cas de « mictionnel », puisque les traductions du français « miction » sont en anglais *micturition* ou *urination*, l'adjectif le plus souvent employé dans les termes français correspondants étant « urinaire »). Les adjectifs pouvant donner lieu au type de variation que nous souhaitons observer sont ici relativement peu nombreux (transfusionnel, nutritionnel, lésionnel, et gestationnel). Nous avons donc décidé de nous concentrer sur l'étude des dix adjectifs relationnels les plus fréquents pour lesquels au moins une occurrence de l'équivalent anglais existait dans le corpus CCAE et pour lesquels le corpus contenait des occurrences de patrons syntaxiques employant le nom et l'adjectif en tant que prémodificateur (soit par exemple *confusional state* et *confusion pattern*).

Le Tableau 3 fait apparaître le nombre d'expressions nominales distinctes formées à partir de noms en *–ion* ou d'adjectifs en *–ional* dans le corpus CCAE (le nombre maximal de résultats affichés par l'interface de recherche du corpus est de 100). Il est difficile de dégager une tendance générale de ces chiffres car la saturation du décompte à 100 peut dissimuler des différences éventuelles entre la prémodification nominale et adjectivale, notamment sur l'ensemble de la partie *ACADEMIC* du corpus pour les formes *nutrition(al)*, *transition(al)*, *organization(al)* et *institution(al)*. Certaines formes nominales (*gestation*, *emotion*, *direction* et *institution*) sont nettement moins employées en médecine que leurs équivalents adjectivaux. Les seuls adjectifs qui soient utilisés exclusivement dans la sous-section médicale de la partie *ACADEMIC* du corpus sont *lesional*, *transfusional* et *confusional*, et la prémodification nominale est nettement majoritaire par rapport à la prémodification adjectivale pour les deux premiers.[10]

[10] Ces emplois semblent marginaux et pourraient provenir de l'usage de scientifiques des pays de langues romanes. Ainsi, les moteurs de recherche de la Toile ne donnent que 4 emplois de *transfusional risk* sur les sites britanniques pour plus de 300 occurrences de *transfusion risk* et 46 emplois de *transfusional safety* pour plus de 2000 occurrences de *transfusion safety*.

	Articles de recherche (ACADEMIC)		Articles de médecine (ACAD:Medicine)	
	Nom Nom	**Adj Nom**	**Nom Nom**	**Adj Nom**
nutrition(al)	100	100	29	46
lesion(al)	28	7	25	7
transfusion(al)	22	1	21	1
gestation(al)	3	14	1	7
transition(al)	100	100	43	12
emotion(al)	46	100	2	93
organization(al)	100	100	42	100
direction(al)	56	100	3	23
institution(al)	100	100	7	67
confusion(al)	22	4	3	3

Tableau 3. Nombre d'expressions nominales distinctes formées à partir de noms en *–ion* ou d'adjectifs en *–ional* dans le corpus CCAE

L'examen des divers noms prémodifiés dans les deux types de structures montre un faible taux de recoupement. Ainsi, il n'y a pas de variation entre prémodification nominale et adjectivale dans les expressions nominales de fréquence supérieure ou égale à 2 dans la partie médicale du corpus dans le cas de l'adjectif *nutrition(al)*, à l'exception du mot *information*.[11]

nutrition manager	8
nutrition information	7
nutrition examination	7
nutrition care	5
nutrition therapy	5
nutrition status	3
nutrition staff	3
nutrition assessment	2
nutrition policies	2
nutrition services	2
nutrition board	2
nutrition intervention	2
nutrition program	2

nutritional status	28
nutritional anaemia	9
nutritional management	5
nutritional support	5
nutritional intake	5
nutritional counseling	4
nutritional requirements	3
nutritional value	3
nutritional treatment	2
nutritional rehabilitation	2
nutritional history	2
nutritional capillaries	2
nutritional education	2
nutritional factors	2
nutritional diseases	2
nutritional deficiencies	2
nutritional information	2

Tableau 4. Fréquence des expressions nominales contenant une prémodification par *nutrition(al)* dans la partie médicale du corpus CCAE

Une comparaison faisant intervenir la prémodification par *incision(al)* montre relativement clairement ce qui distingue les noms prémodifiés sémantiquement parlant. Les noms prémodifiés par l'adjectif indiquent soit la manière dont est effectué un geste diagnostique (la biopsie) soit la cause d'un état patholo-

[11] La majorité des emplois de *nutrition information* se retrouvent dans la formule *NUTRITION INFORMATION # (per serving).*

gique (hernie, douleur, blessure).[12] Ceux qui sont prémodifiés par le nom *incision* ont principalement une fonction localisante (*site*, *line*, *point*).

incisional hernia(s)	15
incisional biops[y/ies]	5
incisional pain	2
incisional wound	1

incision site (s)	5
incision line	2
incision point	1
incision scar	1
incision surgery	1

Tableau 5. Nombre d'expressions nominales contenant une prémodification par *incision(al)* dans la partie médicale du corpus CCAE

La présence d'hapax rend cependant difficile une généralisation de cette règle, et le contexte peut révéler des contraintes syntaxiques qui priment sur un éventuel choix (le contexte de l'expression *incision surgery* révèle ainsi la présence de l'adjectif *small* formant l'expression figée *small incision surgery* : [...] *elevators, dissectors, scissors and punches, all designed for small incision surgery*). Dans le cas du choix possible entre *incision scar* et *incisional scar*, la consultation des moteurs de recherche de la Toile confirme clairement l'usage moindre de l'adjectif relationnel par un facteur 5. Notons que le corpus MED_FR, de taille presque deux fois supérieure au corpus CCAE, ne contient que peu d'occurrences de l'adjectif *incisionnel*, employé sept fois en association avec deux noms (hernie et tracé), dont une structure attributive (Les hernies périnéales peuvent être spontanées ou incisionnelles).

Fréquence	**Nom**	**Adjectif**
25	mécanisme	lésionnel
20	diagnostic	lésionnel
19	œdème	lésionnel
19	niveau	lésionnel
12	profil	lésionnel
9	processus	lésionnel
5	syndrome	lésionnel
4	stade	lésionnel
4	aspect	lésionnel
4	examen	lésionnel
3	état	lésionnel
3	foyer	lésionnel
3	site	lésionnel
3	risque	lésionnel
3	volume	lésionnel
2	phénomène	lésionnel
2	blush	lésionnel

Tableau 6. Expressions nominales de fréquence supérieure ou égale à 2 contenant l'adjectif *lésionnel* dans le corpus MED_FR

L'étude de corpus comparables visant à la recherche d'équivalents de traduction peut s'avérer ardue dans le cas de lexèmes ayant plusieurs équivalents de traduction dont certains n'ont pas de lien étymologique direct avec le terme de départ. Ainsi une recherche des traductions des nombreuses expressions nominales utilisant l'adjectif « lésionnel » dans le corpus MED_FR donne peu de résultats dans le corpus CCAE, où la prémodification par *lesion* ne donne que *lesion level* (niveau lésionnel), *lesion site* (site lésionnel) et *lesion pattern* (profil lésionnel), la prémodification par l'adjectif *lesional* ne donnant que les

[12] La traduction du terme *incisional biopsy* présente dans le Grand Dictionnaire Terminologique (*biopsie d'incision*) est absente du corpus MED_FR, où l'on relève toutefois plusieurs occurrences de l'expression *incision-biopsie*. Les utilisations de *biopsie incisionnelle* sont majoritaires sur la Toile par rapport au total des occurrences de *biopsie d'incision* et *biopsie par incision.*

expressions *lesional skin*, *lesional biopsy*, *lesional T-cells*, *lesional measurements* et *lesional areas* (autre traduction possible de « site lésionnel »). La majorité des expressions du Tableau 6 n'auraient ainsi pas d'équivalents de traduction dans le corpus CCAE.

L'étude dans le corpus CCAE de la prémodification par *injury* (qui signifie lésion ou blessure selon les contextes) permet cependant de détecter un certain nombre d'expressions pouvant constituer des équivalents de traduction de celles du Tableau 6 (*injury mechanism*, *injury rate*, *injury criteria*, *injury severity*, *injury site*, *injury pattern*, *injury region*, *injury incidence*).

6 Conclusion

Les données statistiques tirées des corpus comparables utilisés ici montrent un certain degré de variation, que la présence étude contribue à décrire sans toutefois l'expliquer dans ses détails. Une piste semble cependant pouvoir être exploitée : la détermination de critères sémantiques et discursifs pouvant distinguer les noms anglais susceptibles d'être précédés d'un nom de ceux qui sont précédés d'un adjectif, puisque ces deux ensembles semblent relativement hermétiques, comme on l'a montré dans le cas de l'adjectif *nutritional*.

La taille des corpus utilisés, pourtant conséquente pour un domaine de spécialité, peut s'avérer insuffisante pour des études comparables portant sur des adjectifs d'emploi peu fréquent. Le recours aux moteurs de recherche de la Toile est sans doute appelé à se développer dans ce domaine, en dépit des problèmes méthodologiques qu'il peut poser. Une utilisation fine de ces moteurs de recherche permettant de distinguer, quand cela est possible, les textes produits par des anglophones natifs est en tout cas nécessaire, l'utilisation d'adjectifs comme *lesional* ou *transfusional* semblant avoir pour origine les écrits de chercheurs en médecine qui sont souvent des locuteurs non natifs. L'adjectif relationnel, dont l'emploi est plus courant dans les langues d'origine romane, pourrait également faire l'objet d'études sociolinguistiques visant à déterminer si son utilisation est effectivement plus fréquente dans les articles scientifiques rédigés en anglais par des locuteurs non natifs.

Références

Bourigault, Didier, Christian Jacquemin & Marie-Claude L'Homme, (eds). 2001. *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam.

Daille, Béatrice. 2001. "Qualitative terminology extraction: Identifying relational adjectives" in Bourigault, Didier, Christian Jacquemin, C. & Marie-Claude L'Homme, (eds). *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, pp. 149-166.

Fradin, Bernard. 2008. "Les adjectifs relationnels et la morphologie" in *La Raison Morphologique: hommage à la mémoire de Danielle Corbin*, John Benjamins Publishing Company, Amsterdam, pp. 69-91.

Frantzi, Katerina & Ananiadou, Sophia 1999. *The c-value/ncvalue domain independent method for multiword term extraction.* Journal of Natural Language Processing, 6(3):145–179

L'Homme, Marie-Claude. 2002. « Fonctions lexicales pour représenter les relations entre termes », *Traitement automatique des langues (TAL)* 43(1), pp. 19-41.

L'Homme, Marie-Claude. 2004. "Adjectifs dérivés sémantiques (ADS) dans la structuration des terminologies", In *Actes de Terminologie, ontologie et représentation des connaissances*, Université Jean-Moulin Lyon-3, 22-23 janvier 2004.

Maniez François. 2005. "Identification automatique des adjectifs relationnels : une étude sur corpus" in *De la mesure dans les termes* (éd. H. Béjoint et F. Maniez), Presses Universitaires de Lyon, Lyon, pp. 134-152.

Mélis-Puchulu, Agnès. 1991. « Les adjectifs dénominaux : des adjectifs de "relation" », *Lexique* 10, pp. 33-60.

Monceaux Anne. 1997. « Adjectifs de relation, complémentation et sous classification », *Langages* 126, pp. 39- 59.

Normand Sylvie & Didier Bourigault. 2001. "Analysing adjectives used in a histopathology corpus with NLP tools". *Terminology* 7(2), pp. 155-164.

Sournia Jean-Charles. 1974. *Le langage médical moderne*, Paris, Hachette.

Thoiron Philippe & François Maniez. 2004. « Les groupes nominaux complexes dans le décodage et la traduction en langue de spécialité : quelles ressources lexicales pour l'apprenant en anglais médical ? » in Lino T. ed, *Etudes de Linguistique Appliquée. Revue de didactologie des langues-cultures et de lexiculturologie*, Paris, Didier Érudition, pp. 327-346.

Warren, Beatrice. 1984. *Classifying adjectives* (Gothenburg studies in English: 56). Göteborg: Acta Universitatis Gothoburgensis.

Dictionnaires

Dorland's Pocket Medical Dictionary, édition française bilingue, Masson-Elsevier, 2008.

Grand Dictionnaire Terminologique, Office québecois de la langue française (http://www.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp).

Grand Robert de la langue française (version électronique 2.0), 2005.

Corpus et ressources en ligne

Corpus of Contemporary American English, compilé par Mark Davies, Brigham Young University (http://www.americancorpus.org/).

Science Direct (base de données contenant le texte intégral d'une sélection de 1900 revues de sciences, médecine, techniques, économie, gestion, psychologie et sociologie), Elsevier (http://elsevier.bibliotheque-nomade.univ-lyon2.fr/).

Towards an integrated analysis of aligned texts in terminology: The CREATerminal approach

Elizabeth Marshman
Observatoire de linguistique Sens-Texte /
University of Ottawa
70 Laurier Ave. E. (401)
Ottawa, K1N 6N5
elizabeth.marshman@uottawa.ca

Patricia Van Bolderen
University of Ottawa
70 Laurier Ave. E. (401)
Ottawa, K1N 6N5
trishvanbolderen@gmail.com

Abstract

Parallel, aligned texts are increasingly available to translators and terminologists as the use of tools such as translation memories increases. Such texts constitute excellent resources for extracting terminological information, including term equivalents, occurrences of terminological relations, and terms' co-occurrents. Unfortunately, however, much of this information — which could be stored on term records for use by individuals or language services — remains untapped because we lack an integrated tool to help extract this information from aligned texts and store it in an easily accessible format. In this paper, we describe the methodology and some results of an analysis of a corpus of aligned texts from popularized health Web sites, to illustrate the kinds of useful information that may be extracted. We then suggest one strategy for developing a computer tool that could accelerate and facilitate this work to allow translators and terminologists to make better use of the aligned texts available to them, and to improve access to the terminological information these texts contain. We call this tool the CREATerminal.

1 Introduction

With the advent of approaches to terminology that question some of the field's traditional tenets (e.g. biunivocity, negative attitudes towards terminological variation and neglecting of terms' relations with other lexical units in discourse), new areas in terminography have opened. There is increasing interest in the study of terminological variation, synonymy, and collocations/co-occurrents. Much of this valuable terminological information has nevertheless not yet been widely integrated into the major terminological resources, and users (e.g. translators) who wish to be able to consult this kind of data must generally find it independently and manage it themselves.

We feel that the growing body of parallel, aligned (or easily alignable) texts produced by high translation volumes and the common use of translation memories constitutes an invaluable source of such information for these users. We will describe a methodology by which language professionals may benefit from the knowledge available to them in their own collections of aligned texts (e.g. previous translations, available parallel documentation), and make the case for developing a tool to assist in this work. Many currently available tools assist in different aspects of this process, and translators and terminologists commonly make use of these in their work. However, a gap remains to be filled by a tool that can integrate, accelerate and facilitate the various steps in the process of analyzing aligned texts to identify information translators and terminologists consider important, while still allowing these users to maintain control of the process, to choose the specific information that is pertinent to them, and to remain in their "comfort zone," working with familiar approaches. Somewhat as Kay (1997) did in his description of the translator's amanuensis — though we do not claim to have his insight or foresight — we hope to highlight the need for such a tool. We hope this discussion will bring us a little step closer to its creation.

In Section 2, we will review some of the literature on the need to study variants, synonyms, relations and co-occurrents for complete terminological description. We then describe a methodology for manually processing aligned texts to extract this information for human interpretation and storage in Section 3, and discuss some examples from a small pilot project in Section 4. We then expand on this idea in Section 5, to propose the development of a computer tool to help terminologists, writers, translators and other language professionals working in specialized fields to extract and store pertinent information: the CREATerminal (from *Co-occurrent and Relation Extraction from Aligned Texts for Terminology*). In Section 6, we will discuss some questions in and limitations of this work, before concluding in Section 7.

2 Background and important concepts

It has been observed (e.g. in L'Homme, 2005) that the definition of "term," one of the central concepts in terminology, is a subject of some debate. Analyses carried out from various perspectives have produced many different definitions that are not universally recognized. As L'Homme points out (2005: 1116-1122), differences in perspectives at a theoretical level have very concrete effects on terminographical approaches. Users thus require flexible termbase structures that allow them to store the kinds of information they consider pertinent from their specific perspectives, which may not be available or easily accessible in conventional terminological resources. Some of this information is described below.

While traditional terminology discouraged the use of synonyms, it has been observed (e.g. Gaudin, 1991; Temmerman, 2000; Cabré, 2003) that synonymy is both undeniable and useful in many contexts. Moreover, other kinds of term variants are frequently observed (e.g. L'Homme, 2004:72-75; Daille, 2005) and are pertinent for terminologists and translators. The analysis of synonyms and some variants is thus an important part of terminological description in a specialized field. Users may wish to identify and store many possible term variants or synonyms, treating them either as equally valid terms (though perhaps intended for use in different circumstances) or as a preferred term and a set of discouraged variants.[1] A terminology tool to meet these needs thus must facilitate the identification and storage of variants in both source and target languages.

Whichever terminology theory provides the framework for terminography, terminological relations play an essential role in identifying, defining and describing terms and their meanings or the concepts they denote (Meyer & Mackintosh, 1994; L'Homme & Marshman, 2006; L'Homme, 2005, 2006). However, these relations are unfortunately not widely and explicitly included in terminological databases. Although Meyer and her colleagues (1992) proposed creating a terminological knowledge base (TKB), and this term continually recurs in the literature (e.g. Condamines & Rebeyrolle, 2001), inclusion in mainstream terminological resources (e.g. term banks) is rare. Users who want access to this type of information must generally collect and store it in their own term records. However, term record models in terminology management tools are not always well suited to this task. A terminology tool to meet these needs must thus facilitate the analysis of term occurrences to identify relations that are pertinent for the user, and assist in the storage and classification of examples that can be used as a basis for further analysis. (Cf. L'Homme (2005) for a discussion of some different approaches to this analysis.)

Many researchers, including Pavel (1993), Meyer & Mackintosh (1994), L'Homme & Meynard (1998), Pearson (1998), and Heid (2001; Heid & Freibott 1991), have discussed the role of collocates and co-occurrents in terminology as important components of idiomatic specialized discourse. As observed by Heid (2001:788) and Maniez (2001:557), co-occurrents are often not parallel in different languages. This leads to a clear need to identify, establish equivalences between and store co-occurrents in source and target languages in terminological resources in order to facilitate idiomatic expression in texts and translations. Unfortunately, while some (e.g. Heid & Freibott, 1991; Pavel, 1993) have proposed methodologies for including this information in terminological resources, their description in large-scale resources gen-

[1] Even in the latter case, the choice either to choose and encourage — or impose — the use of a preferred term or to accept multiple terms according to the context or users' judgment can only be made after variation is thoroughly evaluated.

erally leaves much to be desired.[2] Again, users are obliged to gather and store this information independently, and require a terminology tool that will facilitate the identification of collocates and co-occurrents and facilitate the process of storing and classifying them for easy future reference.

Section 3 will outline one method of gathering these types of terminological information and storing it in a structure that is designed to facilitate later use, as a precursor to describing how an integrated tool could help facilitate and accelerate the process.

3 Pilot project methodology

This section will describe the methodology used in our pilot project, including the corpora, the termbase design and the analysis.[3]

3.1 The corpus

The corpus used in this project contains 188 aligned popularized texts taken from bilingual Canadian health Web sites. These are good examples of texts that are translated in significant volumes in Canada, and also were expected (cf. Pearson, 1998) to offer examples of a number of phenomena pertinent for this kind of analysis and methodology development, such as synonymy and explanations of terminological relations. The total size of the corpus was approximately 230,000 tokens in English and 269,000 in French. The texts were aligned with the LogiTerm bitext aligner (Terminotix, 2009).

3.2 Design of the termbase structure

All extracted information must of course be stored in an organized, integrated termbase structure. In this pilot project, the termbase model was designed in Microsoft Access (but intended to be adaptable to other types of tools in future development). The starting point of the termbase structure consists of a selection of fairly standard fields, including source and target language terms, contexts and sources. In this termbase, each context is accompanied by its matching target-language segment to facilitate comparison. These contexts are stored in a table linked to the main structure by the source term. In the pilot project, fields for data such as the domain, client, definitions and administrative information have been omitted; of course in future, larger-scale projects these would likely be added.

The structure (Figure 1) diverges from the standard termbase layout with the inclusion of fields for relation occurrences, co-occurrents and bilingual contexts containing examples of these, along with the sources of these data and a basic set of labels for classifying co-occurrents and relation occurrences to facilitate sorting and consultation.

[2] One exception is the DiCoInfo (L'Homme, 2009), in which co-occurrents of Internet and computing terms are represented in a format that is easy to use and to interpret thanks to a classification based on Lexical Functions (LFs) (Mel'cuk et al. 1995; also used by e.g. Clas (1994) and Fontenelle (1994)), accompanied by more popularized paraphrases of these relationships for users who are not familiar with LFs, in the style of resources such as the *Lexique actif du français* (Polguère, 2003; Mel'cuk & Polguère, 2007).

[3] This approach owes much to the DiCoInfo project (L'Homme, 2008), although it clearly differs in a number of areas (e.g. using aligned texts and the approach to choosing terms to describe, including multi-word terms that are compositional in their meanings).

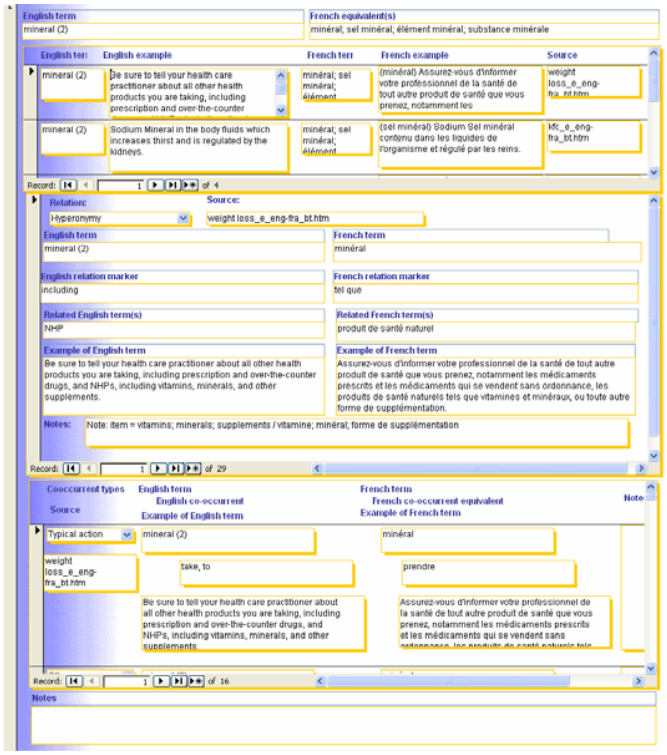


Figure 1. Preliminary term record template

3.3 The analysis

The analysis of the corpus data was carried out in five main steps: term extraction, identification of contexts and equivalents, identification of synonyms and variants, identification and classification of terminological relations, and identification and classification of co-occurrents.

Term extraction from the English corpus texts was done using TermoStat (Drouin, 2003). The first 250 candidates were reviewed and those that were clearly noise or non-terminological units (e.g. *http*, *asp*, *substance*, *type*, *be*) were removed. The remaining 175 candidates were then imported into the termbase.

The next step was the search for useful contexts and equivalents. The LogiTerm bilingual concordancer (Terminotix, 2009) was used to search for all occurrences of the retained terms. We then analyzed the aligned segments to identify the French equivalents used for the term, and entered these in the termbase. We chose one or more pairs of aligned segments to include in the base as contexts, selecting those we felt best illustrated the term's meaning and/or use and the equivalents identified.

As the forms of terms may vary, and synonyms may be used, the equivalents identified for the original terms were used to search the French segments in the bitexts. Any synonyms or variants of the original terms identified were also added to the base.

The next step involved identifying occurrences of important terminological relations for a selection of the candidate terms. For the purposes of this project, the relations chosen were hyperonymy, meronymy, cause-effect, entity-function, and association.[4] Concordances for each English term were generated, and the aligned contexts extracted were evaluated to identify whether one or more of these relations was present. Where a relation was identified, the context was copied to the termbase and annotated to identify the type of relation, the item to which the term was related and the marker of the relation. A similar process

[4] We included all occurrences of the included relations, whether or not they were repetitions (as some sentences recurred in texts from the same source), and regardless of the nature (terminological or other) of the item linked to the original term by the relation marker. Cases in which the identification of the corresponding items in French was impossible (e.g. due to alignment, file format or content problems) were excluded.

was carried out for the corresponding French segment, providing parallel data for inclusion in the base (i.e. the term equivalent used, the equivalent of the linked term, and the French relation marker, if any).

In the final step, the contexts were analyzed again, this time to identify English terms' co-occurrents and to extract them and the contexts in which they occurred, accompanied by their matching segments and translations in French and their source. The co-occurrents were labeled with a very general indication of the relationships between the co-occurrents and the terms (e.g. light verbs, typical actions, intensifiers, attenuators, qualifiers). As other types of co-occurrents were also observed, an "Other" category was added for gathering additional examples for evaluation and possible later classification.

This manual analysis represents a labour-intensive approach to extracting this kind of data from corpora, but one that is likely to be used by working translators and terminologists. It will serve as the starting point for the discussion in Section 5 of improvements that could be made by building a custom tool that can partially automate some tasks. First, however, some of the results of the pilot project will be described to demonstrate the usefulness of this type of work.

4 Results of the pilot project

This section will briefly present some examples from the results of the ongoing pilot study of a set of candidate terms, to illustrate the kind of data that can be extracted from aligned texts using this methodology. This admittedly very small sample nevertheless allows us to see a range of potentially useful information. We will also discuss some observations of the process and the product of the analysis.

4.1 Results

The 175 candidate terms added to the base included a number of relatively specialized terms such as *neonatal hepatitis*, *hemochromatosis* and *peritoneal dialysis*, as well as some less specialized terms that we considered to be important in the domain, e.g. *follow-up care*, *protein* and *red blood cell*.

While half of the 28 terms we have so far evaluated to find equivalents, synonyms and variants showed one-to-one correspondences, several cases of variation and synonymy were observed. For example, for the candidate term *liver transplant*, two potential equivalents (each with two variants) were identified: *transplantation de foie* (*transplantation du foie*) and *greffe de foie* (*greffe du foie*). Moreover, in searching using these terms we also found *liver transplantation* as a potential variant of the English term. In another case, the term *mineral* (in the second of two senses identified in the corpus, shown above in Figure 1) was found to be translated by such terms as *sel minéral*, *élément minéral* and *substance minérale* in addition to *minéral* alone. Such variants, if made available on term records, could assist users in evaluating potential differences in meanings/concepts denoted, finding potential translations for the terms in various contexts, and/or evaluating existing variations and choosing a preferred term and/or equivalent term to increase terminological uniformity in future texts.

For these 28 terms, we found approximately 360 occurrences of our relations (including repetitions): 138 cause-effect, 82 meronymy, 67 hyperonymy, 45 entity-function, and 28 association. Some examples are shown below.[5] These relations often provide key information about how a term and its meaning/the concept it denotes fits into the system in a domain, assist in comprehension and help to create conceptual/lexical structures for domains (e.g. types of dialysis, causes of hepatitis, cancer risk factors).

Relation	English segment	French segment	Source
Hyperonymy	[W]e discuss the two <u>types of</u> *dialysis* which are used to treat the later stage of chronic kidney disease: hemodialysis and **peritoneal dialysis**.	[N]ous expliquons en quoi consistent les deux <u>types de</u> *dialyse* utilisés pour traiter une maladie rénale chronique à ses derniers stades : l'hémodialyse et la **dialyse péritonéale**.	Kidney Foundation of Canada, 2006

[5] The original terms are indicated in bold, the relation markers are underlined, and the related items are in italics.

Meronymy	Many *salt substitutes* contain potassium and other unsafe **minerals**.	De nombreux *substituts de sel* contiennent du potassium et d'autres **minéraux** dangereux pour votre santé.	Kidney Foundation of Canada, 2006
Cause-effect	Infants with **neonatal hepatitis** caused by the *cytomegalovirus*, *rubella* or *viral hepatitis* may transmit the infection to others who come in close contact with them.	Les nourrissons atteints d'**hépatite néonatale** due au *cytomégalovirus*, au *virus de la rubéole* ou aux *virus des hépatites* peuvent transmettre l'infection aux personnes en contact étroit avec eux.	Canadian Liver Foundation, 2008
Entity-function	The enlarged **vein** is then used as the *access site* for inserting needles to connect you to the dialysis machine.	Cette **veine** dilatée sert alors de *site d'insertion* des aiguilles qui vous relient à l'appareil de dialyse.	Kidney Foundation of Canada, 2006
Association	*Smoking* is a major risk factor for oral and dental disease, including **oral cancer**.	L'*usage du tabac* est un facteur de risque important de maladies buccales et dentaires, dont les **cancers buccaux**.	Health Canada, 2008

Table 1. Examples of relation occurrences from the texts

For the 28 terms we have analyzed, over 250 potential co-occurrents were identified, approximately 40 adjectives (qualifiers, intensifiers and attenuators), 120 verbs (light verbs and verbs indicating typical actions), and 90 other items (e.g. nouns). For example, the term *dose* (usually rendered in French as *dose*, more rarely as *quantité* or *niveau*) was observed with a number of qualifiers, including the intensifiers *high* (Fr. *forte*, *élevée*) and *large* (Fr. *forte*), the attenuators *small* (Fr. *faible*, *petite*) and *low* (Fr. *faible*) qualifiers *prescribed* (Fr. *prescrite*) and *recommended* (Fr. *recommandé*) and a number of verbs denoting typical actions such as *give* (Fr. *donner*), *adjust* (Fr. *ajuster*) and *reduce* (Fr. *réduire*). While most of the co-occurrents above (and in fact, most identified in this analysis) are similar in both languages, it is nevertheless necessary to confirm this parallelism rather than to assume that it will hold. For example, variation is observed in the intensifiers and attenuators, with *high* in English corresponding not only to *élevée* (lit. *high*) but also to *forte* (lit. *strong*), and *low* corresponding not to a literal equivalent of this word in French (e.g. *basse*, *peu élevée*, lit. *low*, *not very high*) but rather to the antonym of *forte*, *faible* (lit. *weak*).

4.2 Distinguishing senses

We created separate records for different senses of terms, e.g. for *protein* used to denote a type of molecule (protein (1), as in *these proteins are produced by the cells*) and a type of nutrient (protein (2), as in *you should increase the protein in your diet*). Criteria for differentiation, described e.g. by Meyer & Mackintosh (1994) and L'Homme (2005, 2008), included differences in terminological relationships and co-occurrents, as well as in the use of equivalents and the behaviour of the terms (e.g. the use of *protein* as countable in the first sense described above, and as uncountable in the second).[6]

4.3 Product of the analysis

This approach produced a number of resources: 1) a set of (at least partially) disambiguated English terms, variants, synonyms and French equivalents; 2) a set of bilingual contexts illustrating terms' use and

[6] This process has nevertheless not yet been completed; additional distinctions no doubt remain to be made as more evidence is gathered.

meanings; 3) a set of bilingual terminological relation occurrences involving the terms, classified by relation type and the terms or other linguistic items involved; 4) an annotated list of terms' co-occurrents, with French equivalents and examples of use; and 5) a set of relation markers labeled by relation type, an indication of the terms they were used with, equivalents in French and examples of use. These come together in an enriched term record, and some components may be useful for additional applications. The relation marker list, for example, may be a rich source of data for applications such as (semi-)automated identification of terminological relations (cf. Section 5 below).

4.4 Challenges

A number of challenges were encountered during the work. Some were clearly practical, such as problems with alignment or omissions of segments or of sub-segments in the files analyzed and with reformulations that produce discrepancies in contexts containing items of interest (including some observed by Maniez (2001)). Others involve more fundamental issues (e.g. choosing how to classify relations and/or types of co-occurrents, to distinguish between senses of terms, and to evaluate potential interference). The impact of some practical challenges can be minimized (e.g. by ensuring that the user has access to the full bitexts to deal with alignment problems), and there is a rich body of literature that describes possible strategies for further developing the approach to make it as useful as possible. Moreover, as the user remains in control of the process, judgment may be used at any stage in interpreting results.

5 An application to support similar analyses: The CREATerminal

The methodology and results above establish that each task described can be accomplished with existing tools (although in a quite labour-intensive way). In fact, many commercial translation environments, as well as a number of systems designed for research, offer many of the components required. The goal of the CREATerminal approach, then, will be to integrate these various functions into a workflow that will allow users such as translators, terminologists and other language service providers (and perhaps also terminology trainers and trainees) — users who often have access to aligned texts but lack a tool to facilitate accessing and storing the valuable terminological information they contain — to carry out each task, evaluate and validate the results, and then move smoothly on to the next task, storing the results of each step in a centralized termbase specifically adapted for their storage, while still offering these users options for personalizing the base and the process. This description will begin with the most basic framework and functions of such a tool: the proposed CREATerminal. It will continue with a description of some possible avenues for increasing the automation of some functions. The basic workflow is illustrated in Figure 2.

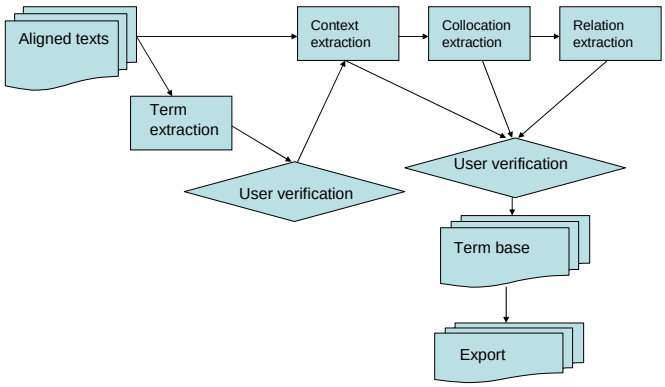


Figure 2. The CREATerminal workflow

Modules

The development of the CREATerminal would require a transition from Access to a more flexible database management system that would allow easy input and output of data and the creation of fields required to store the standard data extracted as well as user-defined fields for additional information (e.g. domain, sub-domain, client/project, definition, administrative information). As well, the tool must be able to import aligned texts in formats produced by various aligners.

In the basic CREATerminal, the termbase structure would store: 1) the main term record, including equivalents, synonyms, variants, and sources; 2) one or more bilingual contexts and their sources; 3) relation occurrences, including terms, equivalents, a relation label, relation markers in both languages, bilingual contexts and sources; 4) co-occurrents, including terms, term equivalents, co-occurrents in both languages, bilingual contexts, and sources; and 5) complete references for the aligned texts processed.

The modules of the CREATerminal would feed information automatically to the appropriate structures after user evaluation and verification, a phase which would be made as quick and easy as possible (e.g. by offering checkboxes for users to select data to send to the termbase), and which would allow access to the complete collection of aligned texts at any point to assist in evaluation.

The process would begin with source-language term extraction, user validation of candidate terms, and the importing of selected terms into the termbase. Users should also be able to input additional terms manually, if desired. In the next stage, an integrated bilingual concordancer would be used to generate queries in the aligned texts (ideally automatically taking into account inflected forms and some typical variants). Users could then scan the aligned segments, highlight potential equivalents, select contexts to be included in the term records, and send them to the termbase (automatically accompanied by the sources of each piece of the information). The equivalents, once identified, would then be used to search the aligned texts for occurrences (e.g. targeting segments that did not contain the original search terms), to highlight potential variants or synonyms. This would produce a very basic term record which of course could be complemented by additional data as the user wished.

The next stage would involve searching bilingual concordances for contexts indicating terminological relations. Users would identify pertinent contexts and highlight both the terms or other linguistic units that participate in the relationship and the relation markers that indicate this relationship, and finally classify the relations observed using one of a set of simple relation labels (or potentially a custom label). The accumulated information (including sources) would then be added automatically to the termbase.

The final stage of the analysis would involve the identification of co-occurrents, again using bilingual concordances generated using the source-language terms. As in previous steps, users would specify the desired bilingual contexts, highlight the co-occurrents identified in each language, and apply a label to identify the type of relationship observed between the term and its co-occurrent. This might be done using very general labels or a more detailed classification. Users ideally would be free to select a set of labels (e.g. based on LFs) or potentially to specify their own. Again, validated information would then be sent to the appropriate data structure, completing the basic CREATerminal process.

It should be possible at any point in this process for users to carry out additional searches to look for information and to add this to the term record (e.g. in user-defined fields) if desired. In addition, as observed in the pilot project, the search for equivalents, co-occurrents and terminological relations may well reveal ambiguities and thus a need to create separate term records for a single form. The user should be able at any time to split a record and assign individual pieces of information to each new record.

At the end of the process, output options should be maximized, to ensure that the data can be exported to a variety of terminology management systems or other programs if desired.

Increasing the level of automation

One of the strengths of this kind of approach is that once the general framework is in place it should permit (but not impose) a gradual increase in the level of automation in the analyses. While the initial implementation of the CREATerminal could include only a terminology database for storing the information described above, a monolingual term extractor, and a bilingual concordancer with pathways linking the functions of each, additional tools could be added to assist human users in most of their tasks.

While users will likely have access to a bitext aligner (either in a translation environment or a stand-alone tool, either free or commercial), it would also be possible to include an aligner in the CREATerminal, to provide seamless integration from original source and target texts to termbase product.

Many researchers (e.g. Névéol and Ozdowska, 2005; Deléger et al., 2006; Kraif, 2008) have evaluated the possibilities of automatically identifying equivalents of terms and other items in large parallel corpora. A number of commercial translation environments already implement functions that allow the extraction of term candidates' equivalents from smaller collections of aligned texts. If sufficient occurrences of a given term are available, it should be possible to apply an automatic approach to propose candidate equivalents to the user for verification, accelerating and facilitating this process considerably.

The (semi-)automatic extraction of relations from electronic texts has a long history in terminology (e.g. Meyer et al., 1999; Condamines & Rebeyrolle, 2001; Malaisé et al., 2005; Barrière & Agbago, 2006; Halskov & Barrière, 2008). Projects have focused on the identification of relations to help in tasks from acquisition of domain knowledge to automatic construction of ontologies.

Once a significant set of relation markers has been collected (or using a starter list), it would be possible to use these markers to search the corpora semi-automatically for relation occurrences. This processing could help rank contexts to propose those most likely to be useful first, and even sort them according to the type of relation likely to be present, thereby accelerating and structuring the analysis. Offering the user the opportunity to choose relations of interest, and even markers of interest, could also help users to target the most useful occurrences.

Moreover, existing techniques for automatically identifying co-occurrents and collocates (e.g. discussed in Heid, 2001:803-806), in order to rapidly generate a list of candidates that could be verified and supplemented by the user, could also accelerate the analysis.

The goal of automation would of course be to support, rather than replace, the user's analysis and evaluation. As such, these functions would be activated on demand, rather than automatically.

6 Questions in and limitations of the approach

A first question about the approach relates to its fundamental validity and concerns the analysis of aligned texts. Concerns are often raised in the field of terminology about the use of translated texts and other resources. These largely focus on the possibility of interference from the source language in a translated text, and a resulting lack of idiomaticity of these texts. To give only one example, Dubuc (2002:51) advises very strongly against the use of translated texts for thematic terminology work.

We believe, however, that recent years have seen an increasingly open attitude towards the use of translated texts (e.g. Botley et al., 2000; Maniez, 2001; Bowker and Pearson, 2002) as resources for translation and terminology. Widely used translation tools such as translation memories are intended to aid in the re-use of segments of translated texts, and are most commonly implemented for translating specialized and technical (and usually terminology-rich) texts. Moreover, research in previously translated texts is often an essential step in determining an organization's/client's preferred terminology, which is essential to maintaining the consistency (if not absolute uniformity) that is the hallmark of good terminology management.

In our opinion, the CREATerminal approach gives users the opportunity to benefit from the rich store of information in translated texts, while still providing an opportunity for these users to review and evaluate their findings in light of complementary research. Users control the texts used, and can thus identify and evaluate the sources of information and their reliability.

Additional strategies may also be implemented to maintain the highest possible quality. First, if appropriate texts translated in both directions are available, it would be possible to search all of these using the CREATerminal approach to obtain balanced results (e.g. as described by Maniez, 2001). Comparable corpora in the target language could be used to research items of interest to confirm and evaluate their use in original texts. (A monolingual concordancing feature could even be added to the CREATerminal framework to assist in this task.)

The CREATerminal approach will also allow users to verify all information extracted at any point, and to add supplementary findings from further research. Of course, this complementary work would require effort on the part of the user, but we argue that this should be no more onerous than the research currently carried out in comparable corpora. The observations from aligned texts would be a valuable starting point that would allow users to quickly build a collection of data that can be verified as desired, and the centralization of this data in an adapted framework would be a considerable benefit to users regardless of whether they carry out research only in aligned texts or in a combination of aligned and comparable texts.

A second point that may spark debate relates to this investment of time and effort in human analysis of texts at a more general level. The inclusion of information not often stored in term records today will necessarily involve a greater investment of time in analyzing and storing data. (However, it is worth noting that the process is considerably faster than the extraction of such information directly from comparable corpora for bilingual work.) The time and effort involved would also certainly be greater than in the case of a more automated approach, although the user's control over the process and the choices made — and thus over the quality of the product — would be considerably greater. In the end, we believe that the benefits of access to this information for guiding the learning and use of terms and associated reductions in costs of revision are likely to offset some of this investment, and that ultimately higher quality (both in the terminology database and in documents using the terminology) will be attained.

The last, but clearest, limitation of this work is of course that the CREATerminal is not yet a reality. Given our little feet, we propose its development in little steps; this discussion is only the first of these. Another will be evaluating the feasibility of combining the kinds of functions identified in other tools into a coherent whole with the requisite level of user-friendliness. This will pose a number of challenges, but ones we hope developers will be willing to rise to meet.

7 Concluding remarks and future work

We hope that we have succeeded in highlighting the possibilities for extracting terminological information from aligned texts. We believe that a tool designed to facilitate such analyses would allow terminologists, translators, translator trainers and other language professionals to quickly and easily gather and store pertinent information. As many existing tools offer one or more of the components we identify as necessary, it seems that similar tools could be combined in a single package to simplify and manage the terminology workflow. This makes us optimistic that it will be possible to design and develop such a tool: the CREATerminal.

Much work remains ahead of us. Further data analysis and development and refinement of the methodology are required to prepare for the development of the CREATerminal framework. It will be essential to expand and refine the classification of co-occurrents and potentially of relations; an analysis of further data will assist in this task. The classification system may, for example, be refined using LFs as a starting point. Further developing and applying the criteria for distinguishing senses of terms should also be a focus in future work. In addition, it would be interesting to evaluate the co-occurrents not only of terms but also of the relation markers identified in the analysis, to assist users in using and translating these items idiomatically. Finally, the analysis of the impact of the difficulties described in Section 6 above and of strategies for solving them must of course continue.

Acknowledgements

The authors thank the Canadian Breast Cancer Foundation, the Canadian Cancer Society, the Canadian Liver Foundation, Health Canada, the Hereditary Breast and Ovarian Cancer Foundation, the Kidney Foundation of Canada and the Trillium Gift of Life Network for consenting to the use of their texts in the corpus. Thanks are extended to Nazila Khalkhali for her work on building and aligning the corpus, the Faculty of Arts and University of Ottawa for financial support, and Lynne Bowker for her suggestions and comments.

References

Barrière, C., & A. Agbago. 2006. TerminoWeb: A Software Environment for Term Study in Rich Contexts. http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48765.pdf. Consulted 18 February 2009.

Botley, S.P., A.M. McEnery, & A. Wilson, eds. 2000. *Multilingual Corpora in Teaching and Research*. Rodopi, Amsterdam/Atlanta.

Bowker, L., & J. Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge, London/New York.

Cabré, M.T. 2003. Theories in terminology: Their description, prescription and explanation. *Terminology*, 9(2):163-199.

Canadian Liver Foundation. 2009. Liver disease. http://www.liver.ca/Liver_Disease/Default.aspx. Consulted 19 February 2009.

Canadian Liver Foundation. 2009. Maladie du foie. http://www.liver.ca/fr/Liver_Disease/Default.aspx. Consulted 19 February 2009.

Clas, A. 1994. Collocations et langues de spécialité. *Méta*, 39(4):576-580.

Condamines, A., & J. Rebeyrolle. 2001. Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CKTB): Method and Results. In D. Bourigault, C. Jacquemin & M.-C. L'Homme, eds. *Recent Advances in Computational Terminology*. 127–148. John Benjamins, Amsterdam/Philadelphia.

Daille, B. 2005. Variations and application-oriented terminology engineering. *Terminology*, 11(1):181-197.

Deléger, L., M. Merkel, & P. Zweigenbaum. 2006. Using word alignment to extend multilingual medical terminologies. In P. Zweigenbaum, S. Schulz, & P. Ruch, eds. *Proceedings of the LREC 2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine*. Genoa, Italy. http://estime.spim.jussieu.fr/~pz/lrec2006/Deleger.pdf. Consulted 27 February 2009.

Drouin, P. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99-115.

Dubuc, R. 2002. *Manuel pratique de terminologie*, 4e édition. Linguatech éditeur, Montreal.

Fontenelle, T. 1994. Towards the construction of a collocational database for translation students. *Méta*, 39(1):47-56.

Gaudin, F. 1991. *Socioterminologie: Une approche sociologique de la terminologie*. De Boeck Duculot, Brussels.

Halskov, J., & C. Barrière. 2008. Web-based extraction of semantic relation instances for terminology work. *Terminology,* 14(1):20-44.

Health Canada. 2008. "The Effects of Oral Health on Overall Health." http://www.hc-sc.gc.ca/hl-vs/iyh-vsv/life-vie/dent-eng.php. Consulted 19 February 2009.

Health Canada. 2008. "Effets de la santé buccodentaire sur l'état de santé général." http://www.hc-sc.gc.ca/hl-vs/iyh-vsv/life-vie/dent-fra.php. Consulted 19 February 2009.

Heid, U. 2001. Collocations in sublanguage texts: Extraction from corpora. In S.E. Wright & G. Budin, eds. *Handbook of Terminology Management*, vol. II. 788-808. John Benjamins, Amsterdam/Philadelphia.

Heid, U., & G. Freibott. 1991. Collocations dans une base de données lexicale et terminologique. *Meta*, 36(1):77-91.

Kay, M. 1997. The Proper Place of Men and Machines in Language Translation. *Machine Translation*, 12(1-2):3-23.

Kidney Foundation of Canada. 2006. Living with kidney disease. http://kidney.ca/files/Kidney/aaCompleteManual.pdf. Consulted 19 February 2009.

Kidney Foundation of Canada. 2006. Vivre à sa façon. http://www.rein.ca/files/Kidney/aVivre_Complet.pdf. Consulted 19 February 2009.

Kraif, O. 2008. Extraction automatique de lexique bilingue : application pour la recherche d'exemples en lexicographie. In F. Maniez, P. Dury, N. Arlin & C. Rougemont, eds. *Corpus et dictionnaires de langues de spécialité. Proceedings of the Journées du CRTT, 2006*. 67-86. Lyon, France.

L'Homme, M.-C., ed. 2009. *Dictionnaire fondamental de l'informatique et de l'Internet*. http://olst.ling.umontreal.ca/dicoinfo/. Consulted 18 February 2009.

L'Homme, M.C. 2008. Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire*, 217:78-103.

L'Homme, M.C. 2006. A look at some Canadian contributions to terminology. In H. Picht, ed. *Modern Approaches to Terminological Theories and Applications*. 55-75. Peter Lang, Bern.

L'Homme, M.-C. 2005. Sur la notion de « terme ». *Meta,* 50(4):1112-1132.

L'Homme, M.-C. 2004. *La terminologie : principes et techniques*. Presses de l'Université de Montréal, Montreal.

L'Homme, M.-C. & I. Meynard. 1998. Le point d'accès aux combinaisons lexicales spécialisées : présentation de deux modèles informatiques. *TTR,* 11(1):199-227.

L'Homme, M.-C., & E. Marshman. 2006. Extracting terminological relationships from specialized corpora. In L. Bowker, ed. *Lexicography, Terminology, Translation: Text-Based Studies in Honour of Ingrid Meyer*. 67–80. University of Ottawa Press, Ottawa.

Malaisé, V., P. Zweigenbaum, & B. Bachimont. 2005. Mining defining contexts to help structuring differential ontologies. *Terminology*, 11(1):21-53.

Maniez, F. 2001. Extraction d'une phraséologie bilingue en langue de spécialité : corpus parallèles et corpus comparables. *Méta*, 46(3):552-563.

Marshman, E., & M.-C. L'Homme. 2006. Disambiguating lexical markers of cause and effect using actantial structures and actant classes. In H. Picht, ed. *Modern Approaches in Terminological Theories and Applications*. 261-285. Peter Lang, Bern.

Mel'čuk I., & A. Polguère. 2007. *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. De Boeck & Larcier, Brussels.

Mel'čuk, I., A. Clas, & A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. AUPELF-UREF/Éditions Duculot, Brussels.

Meyer, I., & K. Mackintosh. 1994. Phraseme analysis and concept analysis: Exploring a symbiotic relationship in the specialized lexicon. In *Proceedings of Euralex 1994*. 339-348. Amsterdam, The Netherlands.

Meyer, I., D. Skuce, L. Bowker, & K. Eck. 1992. Towards a new generation of terminological resources: An experiment in building a terminological knowledge base. In *Proceedings of COLING-92*. 956-960. Nantes, France.

Meyer, I., K. Mackintosh, C. Barrière, & T. Morgan. 1999. Conceptual sampling for terminographical corpus analysis. In *Proceedings of Terminology and Knowledge Engineering TKE '99*. 256-267. Innsbruck, Austria.

Névéol, A., & S. Ozdowska. 2005. Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français. In *Proceedings of EGC'05*. http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=ozdowska&subURL=publis/neveol-ozdowska-egc05.pdf. Consulted 27 February 2009.

Pavel, S. 1993. La phraséologie en langue de spécialité. Méthodologie de consignation dans les vocabulaires terminologiques. *Terminologies nouvelles*, 10:66-82. http://www.btb.termiumplus.gc.ca/didacticiel_tutorial/francais/contributions_sp/1993_phraseologie_f.htm. Consulted 27 February 2009.

Pearson, J. 1998. *Terms in Context*. John Benjamins, Amsterdam/Philadelphia.

Polguère, A. 2003. Collocations et fonctions lexicales : pour un modèle d'apprentissage. In F. Grossmann, & A. Tutin, eds. *Les Collocations. Analyse et traitement*. 117-133. Travaux et Recherches en Linguistique Appliquée E:1. De Werelt, Amsterdam.

Temmerman, R. 2000. *Towards New Ways of Terminological Description: The Sociocognitive Approach*. John Benjamins, Amsterdam/Philadelphia.

Terminotix. 2009. LogiTerm Corporate version. http://www.terminotix.com. Consulted 18 February 2009.

Explanatory Combinatorial Lexicology and NLP Applications in Specialized Discourse

Leo Wanner
ICREA and Universitat Pompeu Fabra
C. Roc Boronat, 138
08018 Barcelona, Spain
leo.wanner@icrea.es

Abstract

Does computational processing of specialized discourse require a special theoretical framework? For lexical aspects of specialized discourse, we answer this question in the negative. We claim that the lexical issues faced in specialized discourse are the same as in general discourse. And, therefore, in order to successfully deal with them, the same theoretical means are required as for general discourse. This obviously implies that these means must be suitable and exhaustive. The theoretical means we present are those of the Explanatory Combinatorial Lexicology (ECL). Focusing on some selected computational applications in which lexicon plays a particular role, we describe how ECL has been used within these applications.

1 Introduction

Does computational processing of specialized discourse require a theoretical framework different from frameworks used to handle general discourse? Not a long time ago, this question would have been answered in the affirmative without hesitation by both scholars working in specialized discourse (cf., e.g., Budin and Felber, 1989; Cabré, 1992, 2003, 2007; Temmermann, 2000) and scholars in Natural Language Processing (NLP), where specialized discourse applications were worked on under the heading of *sublanguage* (cf., e.g., Kittredge and Lehrberger, 1982; Sager and Nhàn, 2002). Nowadays, this is less clear. An increasing number of works in specialized discourse draw upon general linguistic theories (Shinmori et al., 2003; Dolbey et al., 2006; L'Homme, 2007; da Cunha et al., 2007). On the other hand, general linguistic theories are increasingly judged with respect to their capacity to describe a large variety of linguistic phenomena – no matter whether they stem from general or specialized discourse. This new tendency appeared not by chance: while specialized discourse may reveal specific discourse style (e.g., predefined discourse patterns in stock market bulletins), specific linguistic constructions or specific vocabulary (e.g., telegraphic style in meteorology; jargon with abbreviations and Latin terms in medicine), or require specific pre-processing stages (as, e.g., the stage of interpretation in the case of turbine monitoring numeric time-series which a report generator may receive as input), the kernel of the (computational) linguistic tasks related to specialized discourse are of the same linguistic nature as in general discourse. This is also and particularly true for lexical issues. As demonstrated by Cohen (1986), Frawley (1986), Heid and Freibott (1991), L'Homme (1998, 2007), Daille (2003), Wanner et al. (2005) and others, as in general discourse, in specialized discourse one faces the need to represent (more or less transparent) multiword expressions and collocations, provide definitions of terminological units (which in their nature are not different from the definitions of lexical units in general language), define the subcategorization frames of predicative units, etc. In the light of these works, we might ask whether it is still justified to oppose terminology to lexicology or even to linguistics in general.[2] While it would go far

[2] Obviously, this concerns only the language part of terminology, not the part that is concerned with the organization of knowledge in the individual domains.

beyond the scope of this short paper to engage into the discussion of this potentially polemic question, we agree with Kageura and L'Homme (2008) that the *raisons d'être* of terminology as a field of study are numerous. However, the question concerning a theoretical framework suitable to handle the prominent lexical issues within both general and specialized discourse computational applications remains. In what follows, we argue that Explanatory Combinatorial Lexicology, or ECL, (Mel'čuk et al., 2005; Mel'čuk, 2006) is a promising candidate. We will attempt to show that ECL offers the needed theoretical means, that it is formal enough to serve as a blueprint for implementation, and that it benefits from being an integral part of a comprehensive linguistic framework, the Meaning-Text Theory (Mel'čuk, 1988, 1997).

In the next section, we introduce the central features of ECL on which we will draw in the course of the following sections. Section 3 analyses why ECL is considered to be suitable for computational applications in specialized discourse. Section 4 briefly addresses the topic of the compilation of specialized discourse ECL-type dictionaries, and then Section 5 discusses some exemplary NLP applications in which ECL has been used. Section 6 summarizes our argumentation on specialized discourse, ECL and NLP.

2 Central Features of Explanatory Combinatorial Lexicology

ECL has a number of features that are of relevance to NLP (and thus also to computational applications in specialized discourse). The two most central of them are that (i) ECL offers the formal means for the exhaustive description of the lexical phenomena that are crucial for NLP-applications, and (ii) ECL is fully integrated into a grammatical framework, namely the framework of the Meaning-Text Theory (MTT). To give the reader a flavour of the capacity of the ECL to describe lexical phenomena, we focus on two notions which are characteristic of ECL and which are indispensable for a number of applications in specialized discourse: the notion of government pattern (GP) and the notion of lexical function (LF). The notion of lexicographic definition that occupies a prominent place in lexicography is still less pertinent in NLP since the decomposition of the meaning of the LUs in the way it is suitable to a human

2.1 Government Patterns

Government Patterns (GPs) are frames that specify the projection of the semantic valency structure of a lexical unit (LU) onto its syntactic valency structure – including the surface realization of the latter. A GP thus subsumes what is usually referred to as *subcategorization frame*, but goes beyond it in that it also covers the relation between the semantic and syntactic valencies. In ECL, GPs are traditionally given in terms of a table, with the headline of the table specifying the correspondence between the semantic and deep-syntactic arguments (or *actants*) of the LU in question and the columns providing the surface realization of the deep-syntactic actants, i.e., their subcategorization information. Consider as example the GP of ACCUSATION in Table 1.[3]

Table 1: Government pattern of ACCUSATION

X ⇔ I who accuses	Y⇔ II who is being accused	Z⇔III of what someone is being accused	W ⇔ IV in front of whom someone is being accused
by N N_{poss}	against N of N	of N that S(entence) S(entence)	in front of N in N

The table illustrates that ACCUSATION possesses four semantic actants, which correspond in an isomorphic way with their deep-syntactic equivalents. Each deep-syntactic actant can be expressed in

[3] The GP displayed in Table 1 is not necessarily complete in the sense that some of the deep-syntactic actants may possess further surface realizations not mentioned in the table.

more than one way; cf.: *the accusation by the prosecutor* vs. *the prosecutor's accusation*, *the accusation against John* vs. *the accusation of John* [*lacks any evidence*], *the accusation of murder* vs. *the accusation that she killed her husband*, etc.

As can be seen in the ECL dictionaries (see, e.g., Mel'čuk et al., 1984-1998), an LU can have more than one GP and the correspondence between the semantic and deep-syntactic valency structures is not always isomorphic.

2.2 Lexical Relations and Lexical Functions as a Means to Represent Them

Lexical Functions (LFs) are a formal means for the representation of binary lexical co-occurrence and syntactic and lexico-semantic derivation (Mel'čuk, 1996, 2006). A binary lexical co-occurrence (or collocation) is a combination of lexemes L_1+L_2 such that one of the lexemes (the *base*), keeps its semantics as isolated item and is freely chosen by the Speaker, while the semantics and the selection of the other lexeme (the *collocate*) depends on the base. For instance, in *make* [*a*] *goal* (in the context of soccer) *make* is the collocate and *goal* is the base: *goal* selects *make*, and it is in combination with *goal* that *make* means '[to] score' (and not, for instance, '[to] perform' as in *make* [*a*] *bow* and not '[to] utter' as in *make* [*a*] *statement*).

Syntactic and lexico-semantic derivation is well-known in many grammatical frameworks – although in most of them it has not been formalized. Syntactic derivation includes nominalization (e.g., $move_V$ – $move_N$, *operate – operation*, $strike_V$ – $blow_N$), adjectival derivation (e.g., *luck – lucky*, *happiness – happy*, *sun – solar*), etc. Lexico-semantic derivation includes derivation of the names of the participants of a situation (e.g., *drive – driver*, *drive – vehicle*, *operation – surgeon*, *goal – shot/scorer*), names of the location and time of an act (e.g., *operation – operation theatre*, *game – play time*), name of the set of entities (e.g., *ship – fleet*, *sheep – flock*, *cow – herd*), and so on.

LFs constitute a collocation and lexico-semantic derivation typology such that each LF is an abstract lexico-semantic relation (type) r_i between the base and the collocate or between the base and the derivative, respectively. As a rule, LFs are specified in a functional notation: r_k(B) = C, r_l(H) = D. As type labels, Latin abbreviations are used: Oper_1 for 'perform/do/carry out', Oper_2 for 'undergo', Real_1 in connection with functional objects 'use' and in connection with abstract situations 'act as expected in the situation?, Magn for 'intense/many', S_1 for 'actor', S_2 for patient, S_{instr} for instrument of an action, etc.[4] In total, between 50 and 60 "simple standard" LFs, which capture all high frequency typological variants of collocation and derivations, are distinguished. In addition, complex LFs and LF-configurations are available that are composed out of the simple LFs; cf. (Kahane and Polguère, 2001) for a mathematical calculus for LF composition.

2.3 ECL as a Component of the MTT Model

As pointed out above, one of the crucial advantages of ECL for NLP is that it forms an integral part of a comprehensive linguistic framework, the MTT: the lexical information specified in ECL-dictionaries has a natural place in specific places of MTT's linguistic model that are assigned to it by the theory. In

[4] Two remarks are in order here. The first remark concerns the notation of LFs: in the case of lexical co-occurrence, the subscript specifies the (partial) projection of the actant structure of the base onto its grammatical functions (thus, '1' in Oper_1 and Real_1 means that the first actant of the base is realized as subject and the base itself as direct object, '2' in Oper_2 means that the second actant of the base is realized as subject and the base itself as direct object, etc.); in the case of derivation, the subscript specifies, roughly speaking, the actant concerned (thus, '1' in the case of S_1 means that it is the name of the first actant which is specified, in the case of S_2 that it is the second ,etc.). This syntactic information is extremely important during the use of LFs in computational applications.
The second remark concerns the mathematical nature of LFs: Given that we deal here with 1:*n* relations (a base can have more than one collocate and a head more than one derivative with the same meaning: Oper1(*operation*) = *carry out*, *perform*; Oper_2(*operation*) = *have*, *undergo*; S_1(*hospital*) = *doctor*, *physician*, *medical practitioner*; S_2(*hospital*) = *patient*, *sick person*), the functional notation as introduced above is a simplification. Mathematically, LFs can be interpreted either as maps or as functions that map a single element onto an element set.

order to better understand what this means (and facilitate the presentation of the computational applications below), let us very briefly introduce the MTT model.

The linguistic model of the MTT is a multilevel transduction model; each level is characterized by a distinct level of abstraction at which linguistic structures are represented,[5] and the structures of the same utterance at adjacent levels can be converted into each other using correspondence statements.

In the context of our presentation, three levels are of relevance: semantic (Sem), deep-syntactic (DSynt), and surface-syntactic (SSynt). Sem-structures (SemSs) are predicate-argument networks in which the nodes stand for semantemes and the arcs for the predicate-argument relations between semantemes. DSynt-structures (DSyntSs) are dependency trees in which the nodes are labeled by "deep" LUs and the arcs by deep-syntactic relations: actantial (I, II, III, …), attributive (ATTR), appenditive (APPEND), and coordinative (COORD). The set of deep LUs of a language $\mathcal{L}$ contains all LUs of $\mathcal{L}$ – with some additions and exclusions. Added are two types of "artificial" LUs: (i) symbols of LFs, and fictitious LUs which represent idiosyncratic meaning-bearing syntactic constructions in $\mathcal{L}$. Excluded are: (i) structural words, (ii) substitute pronouns and values of LFs. SSynt-structures are dependency trees with nodes labeled by any type of lexemes (including closed class lexemes) and arcs labeled by syntactic functions (subject, object, …). Figure 1 illustrates the three structures.

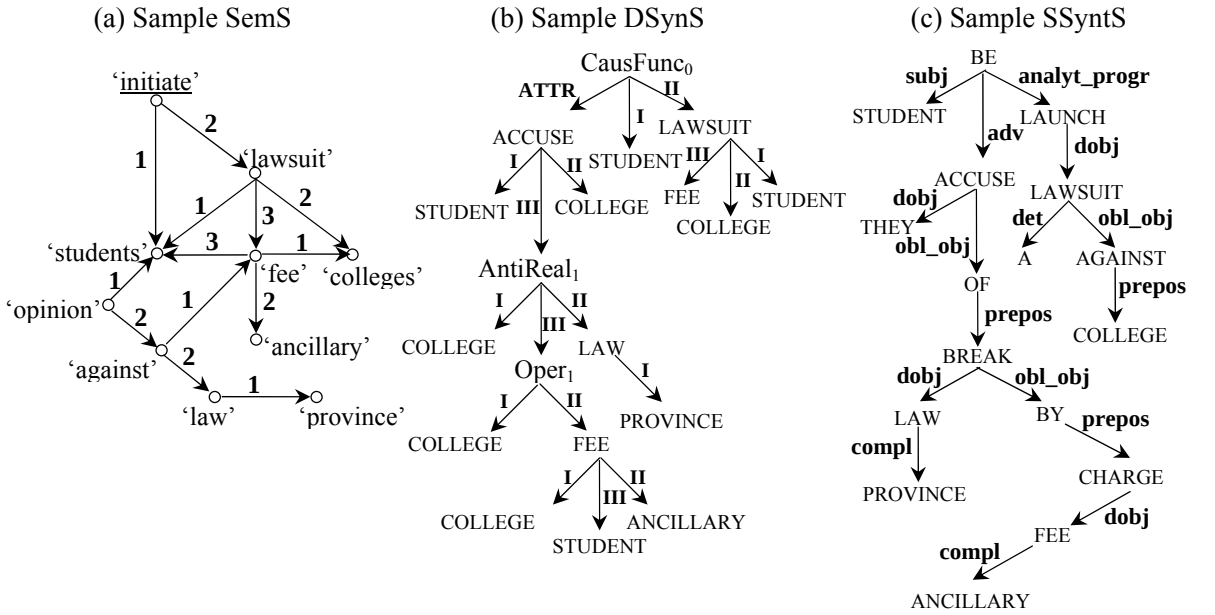


Figure 1: Examples of (simplified) structures in the MTT linguistic model.

The structures in Figure 1 are semantically equivalent in that all three of them represent the sentence *Students are launching a lawsuit against colleges, accusing them of breaking provincial law by charging ancillary fees*.[6] In particular, we see that the semanteme 'lawsuit' corresponds to the lexeme LAWSUIT and

[5] The following presentation is a simplification. In reality, the linguistic representation at each stratum consists of a set of different structures rather than only of one – for instance, the semantic representation counts, along with the semantic structure, the semantic communicative structure and the rhetorical structure. However, for our purposes, such a simplified representation suffices.

[6] The node 'initiate' in the SemS is underlined because it is the "dominant semantic node", i.e., the entry node from which the SemS starts. Note that due to the lack of space, the structures have been simplified. Thus, we skipped in the DSyntS and SSyntS the grammemes attached to the individual nodes (such as number and definiteness for nouns and tense, mode, and aspect for verbs). For the discussion of an older representative list of surface-syntactic

that their actant structures are isomorphic – information that is recorded in the GP of LAWSUIT. The same applies to the correspondence between 'fee' and FEE and 'law' and LAW. 'Initiate' corresponds in the context of 'lawsuit' to the LF $CausFunc_0$. 'Against' −2→ 'law' corresponds to the LF $AntiReal_1$ −II→ LAW; the value of $AntiReal_1$ is stored in the entry for LAW, as the value of $Oper_1$ related to FEE ($Oper_1$ −II→ FEE) is stored in the entry for FEE, and the value of $CausFunc_0$ is stored in the entry for LAWSUIT. The SSyntS reflects the subcategorization information specified in the GPs of the corresponding lexical items (such as the preposition *against* in connection with the realization of the second deep-syntactic actant of LAWSUIT).

3 Why is ECL Appropriate for Computational Applications in Specialized Discourse?

In order to be suitable as a linguistic basis for computational applications in specialized discourse, a lexical theory must fulfill the following two complex conditions:

(i) provide the theoretical and practical means to model all relevant lexical phenomena and to allow for the relation of terms to their conceptual equivalents in order to support applications that start from an abstract knowledge representation;
(ii) be embedded into a comprehensive formal grammatical framework in order to support applications that go beyond individual terms and facilitate a straightforward implementation.

Let us assess how ECL copes with these conditions.

3.1 ECL and Lexical Phenomena in Specialized Discourse

Lexical phenomena that are of particular importance in specialized discourse concern the appropriate description of LUs (or terms) – be they single word or multiword terms – and their combinatorics. Appropriate description means, first of all: (a) lexicographic meaning decomposition to an extent that is useful for the specific application and (b) listing of semantically related terms in a way that these relations can be made use of. Combinatorics means: (a) codification of (the types of) collocations and (b) codification of the actant structure (if applicable).

Meaning decomposition is not common in general discourse computational applications. This is mainly because, on the one hand, hardly any application deals with semantics of a granularity obtained after decomposition of general discourse LUs, and, on the other hand, it is only seldom that a lexicographic definition of a general discourse LU can serve as a paraphrase of this LU, i.e., can be substituted for this LU in an utterance. This is different for specialized discourse multiword LUs. Compositional multiword LUs are one of the main characteristics of specialized discourse in many fields – among them in medicine: *acellular vaccine*, *G6PD deficiency*, *manic depression*, *salivary gland*, etc. A partial decomposition of the meaning of many of these LUs still sounds fine in a sentence when it replaces the LU; cf.: B*ayer experiments with a new acellular vaccine* vs. *Bayer experiments with a new vaccine that does not contain complete cells*, *Mary suffers from an inflammation of the salivary gland* vs. *Mary suffers of the inflammation of the gland that produces saliva*, *John suffers from a manic depression* vs. *John suffers from abnormal highs and downs of mood*, etc. In ECL, lexicographic definitions can be naturally represented in terms of semantic networks; cf. the SemS in Figure 1 above.

Paradigmatic semantic, and in particular taxonomic and meronymic, relations, are central in terminology. ECL offers a number of LFs to model these relations (Gener, $Syn_{\subset}$, $Syn_{\supset}$, Sing, Mult, Anti); for instance, Daille (2003) uses these LFs. However, as pointed out by Fontenelle (1997) and L'Homme (2007:190ff) – and as is natural given the nature of the LFs – the LFs do not cover purely semantic relations that are not restricted lexically. In the past, two different options on how to account for new relations in the ECL-framework have been proposed. The first option is to extend the set of LFs by new

functions of English; see (Mel'cuk and Pertsov, 1987); a comprehensive list of up-to-date Spanish surface-syntactic functions is given in (Mille et al., 2009).

LFs; cf., for instance, (Percova, 1996) for the introduction of a number of derivation LFs (such as Atten for "attenuative", Desc for "descendant", and Fem for "female") and (Grimes, 1990) for the introduction of so-called inverse LFs Specific (as opposed to the standard LF Gener), Whole and Part, Antecedent and Consequent, etc. When choosing this option, one should be aware that the nature of new LFs might (but not need to) be different from the lexico-idiosyncratic nature of the original LFs, which justifies their interpretation as deep LUs. The second option is to introduce a new layer of LFs – as, e.g., Janssen (2005) does with the introduction of a layer of "morphological" LFs and as can be thought of with respect to purely "semantic" LFs. In this case, the different nature of the LF layers does not lead to theoretical clashes.

As far as the collocational part of combinatorics is concerned, Cohen (1986), Heid and Freibott (1991), Laporte and L'Homme (1997), Binon et al. 2000), L'Homme (2007), Vidal (2007) and others convincingly argued that the notion of collocation is crucial in terminology. While in general discourse, the capacity of a system to be able to handle collocations (and thus LFs) tends to provide the system an additional flexibility and power of expression (because other ways to express the same meaning are available), in specialized discourse, collocations often are "the way it is being spoken" in the domain in question.

Consider some examples for LFs from different fields of specialized discourse in Table 2.[7]

The codification of the actant structure in ECL is done by GPs. As we saw above, GPs cover both projection between the semantic and syntactic actant structures of an LU and the subcategorization combinatorics. GPs are used in a number of computational applications – for instance, paraphrasing, text generation and machine translation, no matter whether this LU is of terminological nature or belongs to general language vocabulary: any predicative LU needs to be assigned a GP.

Table 2: Examples of LFs from five different fields

Stock market	$Oper_1$(low/high) = hit; IncepPredPlus(share) = rise; $Real_1$(profit) = lock in; $Real_1$(gain) = cash in; $FinFact_0$(share) = close; $FinFunc_1$(fall) = recover [from]; $AntiReal_1$(market) = overbuy; IncepPredMinus(share) = fall
Computing (partly from L'Homme, 2007)[8]	$Real_1$(program) = run; $CausFunc_0$(file) = create; $CausFact_0$(application) = launch; $PreparFunc_0$(application) = install; $PreparFact_0$(computer) = boot; S_0(compatible) = compatibility; S_1(program) = programmer; S_3(install) = disk; Mult(program) = library; S_{loc}(file) = folder
Medicine	$Oper_1$(diagnosis) = make; $Oper_2$(medical) = have; $IncepOper_1$(illness) = contract; $Real_1$(disease) = succumb [to a ~]; $CausFunc_2$(vaccination) = give; $IncepFunc_1$(virus) = [to] affect; $IncepFunc_0$(epidemics) = break out; AntiVer(health) = poor
Environment	$Oper_1$(temperature) = have; $Func_0$(wind) = blow; $Func_2$(concentration) = be; $IncepFunc_1$(value/threshold) = reach; CausIncepPredMinus(pollution) = cut; AntiMagn(concentration) = low; Magn(heat) = fierce; AntiMagn(heat) = gentle;
Law	$Oper_1$(lawsuit) = pursue; $LiquFunc_0$(lawsuit) = drop; $Real_3$(plea) = accept; $Real_1$(law) = apply; $CausFunc_0$(lawsuit) = launch; $Caus_1Func_0$(law) = pass; $CausFunc_0$(evidence) = produce; S_1(crime) = criminal; S_2(crime) = victim; S_{loc}(custody) = prison; Mult(law) = legislation

When assessing the role of ECL in specialized discourse, we also need to analyze how ECL copes with the necessity to facilitate a bridge between LUs (or terms) and extra-linguistic conceptual units. Numerous works in terminology emphasize the prominence of the notion of concept. As a lexical

[7] Due to the lack of space, we are not in the position to provide a definition of each LF cited in Table 1. The reader is asked to take them as illustrations of the richness of the LF repertoire.

[8] See also DiCoInfo (http://olst.ling.umontreal.ca/dicoinfo/.

framework, ECL does not handle this notion itself. But its modularity on the one hand and the instruments of MTT on the other hand allow for a seamless incorporation of conceptual representations into ECL. Thus, in the same way as a correspondence holds between a semantic and a lexical configuration, a correspondence can be defined between a conceptual and a semantic configuration. And in the same way as traditional GPs model the projection between the semantic valency structure and syntactic valency structure, a similar construct can capture the projection between a conceptual structure and a semantic valency structure. Such a procedure is standard in MTT-based text generators that receive as input an abstract (conceptual) knowledge representation (see below and Lareau and Wanner, 2007).

3.2 ECL as Part of the MTT in Specialized Discourse

As already pointed out above, any NLP application that deals with analysis or synthesis of discourse (no matter whether it is specialized or general) needs to have access to both lexical and grammatical resources. This is true for analysis (parsing) proper and text generation as well as for summarization, machine translation, question answering and even document search. If restricted to a sublanguage, all of these applications can be (and, as a matter of fact, are) viewed as specialized discourse applications; see Section 5 below.

It is thus of primary relevance that the lexical and grammatical models are compatible. As shown in Section 2, this is guaranteed in the case of ECL and MTT since ECL is an integral part of MTT.

Apart from being comprehensive, ECL and MTT are in their nature formal. Cf., e.g., the mathematical calculus for the composition of LFs (Kahane and Polguère, 2001) and the interpretation of the MTT model as a graph grammar model (Bohnet and Wanner, 2001). Formalization is crucial not only for implementation, but also for verification of the wellformedness of the lexical and grammatical resources obtained – which is indispensable when large scale resources are compiled.

4 Compiling ECL Dictionaries for Specialized Discourse NLP

A number of general discourse ECL-type dictionaries for human use are available for French (Mel'čuk et al. 1984–1998; Mel'čuk and Polguère, 2007), Russian (Mel'čuk and Zholkovsky, 1984) and Spanish (Alonso Ramos, 2005). For computational use, small domain-specific ECL-dictionaries have been developed along with the applications for meteorology, air quality, mechanical engineering, optical recording devices, labor market, retail statistics, etc. In general, the computational ECL-dictionaries are application-oriented and thus contain only information used by the application in question. This implies that none of them contains a GP example zone or the lexicographic definition zone – as the human user dictionaries do.

A recent line of research in the context of ECL targets the automatic acquisition of lexical information – first of all of LFs (Daille, 2003; Wanner et al., 2005; Claveau and L'Homme, 2006) and GPs (Mille et al. submitted). Most of this work is situated in specialized discourse – which is, once again, comprehensible given that in specialized discourse the restricted context suggests higher prospects of success.

5 NLP Applications in Specialized Discourse

A number of NLP applications require ECL-type lexical information – among them text mining, opinion mining, information extraction, content distillation, text analysis, summarization, text generation, and machine translation. All of these applications can be also seen in the light of specialized discourse. Let us illustrate the use of ECL in the context of two of them: report generation (as the domain-specific variant of text generation) and machine translation.

5.1 Report Generation

Report generation (RG) is a popular domain-specific (i.e., specialized discourse) application of Natural Language Generation, NLG (Wanner, in print). Report generators (some of them commercial) exist for a number of areas, among them: meteorology (Goldberg et al., 1994; Coch, 1998), stock market exchange

(Kukich, 1983), labor market (Rösner, 1986; Iordanskaja et al., 1992), air quality (Busemann and Horacek, 1997; Wanner et al., 2007), medical care (Bontcheva and Wilks, 2004; Poitet, 2009).

As a rule, RG starts from a numerical time series or a content structure with the goal to produce a written language summary of the most pertinent aspects of the state of affairs observed in the input structure (the criteria of pertinence depend on the preferences of the user and domain-specific restrictions and, in some cases, on the evaluation of a system internal data or knowledge base). If it is a numerical time series, relevant parts of it tend to be mapped onto a concept or a more specific linguistic structure prior to the start of the actual generator.

The majority of the report generators that passed the stage of a prototypical implementation are based on MTT – and thus also on ECL. This is not by chance. Rather, this is due to the features of the ECL and MTT discussed in Sections 2 and 3: availability of GPs (and analogous constructs for the projection of conceptual argument structures onto semantic argument structures), availability of LFs to capture domain-specific idiosyncratic wordings, integration into a multistratal linguistic model that allows for a successive concretization of a linguistic representation up to the surface.

To facilitate an explicit codification and uniform access to all types of ECL-relevant information, in the MARQUIS generator, which produces multilingual air quality information (Wanner et al., 2007), three dictionaries have been defined: a conceptual dictionary, a semantic dictionary and a lexical dictionary (Lareau and Wanner, 2007). The conceptual dictionary is used, first of all, to encode the concept-semanteme mapping information; cf. the simplified entry for the concept of concentration:

```
concentration: property_attribute {
sem = 'concentration'
MATR = {relation = 1 target=referent}
VAL = {relation = 2 target=referent}
ATTR = {relation = 1 source=referent}}
```

The concept CONCENTRATION has two argument slots: something which has a concentration (referred to as MATR in accordance with Sowa (2000)), and a value (referred to as VAL), i.e., an absolute concentration figure. The concept may also be modified by a qualitative characterization of the concentration ("high", "low", etc.), referred to as ATTR. The corresponding semanteme 'concentration' takes MATR as its first semantic argument (indicated by the "relation=1" parameter embedded in MATR's value) and VAL as its second. The attributes "target=referent" and "source=referent" indicate the direction of the semantic relation (for MATR and VAL, the semantic predicate is 'concentration', which takes MATR's and VAL's corresponding semantemes as its arguments, while ATTR's semantic correspondent is a predicate taking 'concentration' as its argument).

The semantic dictionary gives, for every semanteme described, all its possible lexicalisations. For instance, the meaning 'cause' would be mapped to the LUs CAUSE[V], CAUSE[N], RESULT[V], RESULT[N], DUE, BECAUSE, CONSEQUENCE, etc. Note that we do not consider at this stage the valency of the LUs. Thus, it does not matter that *X causes Y* means that *Y results from X*; what interests us here is only that these two lexemes can both be used to denote the same situation, regardless of the communicative orientation. Cf., for illustration, the entry for the semanteme 'concentration' as specified in the semantic dictionary:

```
concentration {
label = parameter
lex = concentration }
```

The semantic type of a semanteme can be specified in the semantic dictionary (cf. "label=parameter"). It is also possible to specify the semantic type of the arguments of a predicate. For instance, adding the attribute "1=substance" here would force the first semantic argument of 'concentration' to be of type "substance".

The lexical dictionary contains, for each LU, all relevant information – including GPs, and LFs; cf.:

```
CONCENTRATION {
     dpos=N // deep part of speech is Noun
     spos=common_noun // surface part of speech is common noun
     person=3 // causes third person agreement
     GP= {   1=I  // first semantic actant is first deep syntactic actant
             2=II  // second semantic actant is second deep syntactic actant
            // First syntactic actant can be realized as ``ozone concentration":
            I={dpos=N // actant is a noun
                rel=compound // linked with compound relation\\
                det=no} // takes no determiner
            // First syntactic actant can be realized as ``concentration of ozone"
            I={ dpos=N  // actant is a noun
                rel=noun_completive // linked with noun_completive relation
                prep=to // takes preposition "to"
                det=no} // takes no determiner
            // Second syntactic actant can be realized as "concentration of 180 \mg":
            II={ dpos=Num  // actant is a number
                  rel=noun_completive // linked with noun_completive relation
                  prep=of // takes preposition "of"
                }}
   Magn = high
   AntiMagn = low
   Adv1 = in // "(we found) ozone in a concentration (of 180 \mg)"
   Func2 = be // "the concentration (of ozone) is 180 \mg"
   Oper1 = have  // "ozone has a concentration (of 180 \mg)"
   IncepFunc2 = reach // "the concentration (of ozone) reached 180 \mg"
   IncepOper1 = reach // "ozone will reach a concentration (of 180 \mg)"
}
```

See (Lareau and Wanner, 2007) for details on the use of these dictionaries in the process of generation.

5.2 Machine Translation

Machine Translation (MT) is one of the most prominent and mature NLP applications – which not always means high quality. It is common sense that the best quality can be achieved when MT is restricted to one specific domain such that the vocabulary and the linguistic constructions can be better controlled. Surprisingly, until recently only a few MT-systems have been developed for specialized discourse; see (Gough and Way, 2003) for an elaboration on this topic. However, no matter whether MT is considered from the domain-specific or general discourse angle, such lexical information as GPs and LFs captured in ECL is essential for both the interlingua and the transfer-based paradigms of MT. Thus, in the case of interlingua-based MT, we face the same needs as in text generation discussed above. In the case of transfer-based MT, if the transfer takes place between surface-oriented syntactic structures, it must account for all syntactic and lexical idiosyncrasies captured by the GPs (in particular the subcategorization part of GPs) and LFs – which leads to a very complex transfer procedure (be it statistical, example- or rule-based). Semantic transfer is not yet feasible. Therefore, Apresjan et al. (1992) and Mel'čuk and Wanner (2001, 2006) argue that the transfer should be done between DSyntSs. This implies that GPs and LFs are made use of during source language parsing and target language generation. Consider, for illustration, the English sentence from above (which we repeat here for the convenience of the reader) and its German translation.

(1) *Students are launching a lawsuit against colleges, accusing them of breaking provincial law by charging ancillary fees*.

(2) *Studenten strengen gegen die Hochschulen, die sie beschuldigen, durch die Erhebung von Zusatzgebühren das Provinzgesetz zu verletzen, ein Verfahren an*
lit. 'Students strain against the colleges, which they accuse, through raising of ancillary.fees, to hurt province.law', a lawsuit

At the first glance their syntactic structures differ significantly. Note also the diverging subcategorization of the translation equivalents ACCUSE and BESCHULDIGEN 'accuse' and BREAK and VERLETZEN 'hurt'. Let us now have a look at corresponding fragments of their DSyntSs in Figure 2 below. These fragments are isomorphic, although their surface realizations are far from being so. The differences are "absorbed", on the one hand, by the GPs of ACCUSE and BESCHULDIGEN, respectively, and, on the other hand, by the value of the LF AntiReal1 in the entry for LAW and GESETZ 'law', respectively:

```
LAW
  AntiReal1: break {prep=by dpos=V form=ger}
GESETZ
  AntiReal1: verletzen {prep=durch dpos=N case=acc}
```

(a) Fragment of the English DSyntS from Figure 1 (b) German equivalent of the English DSyntS in (a)

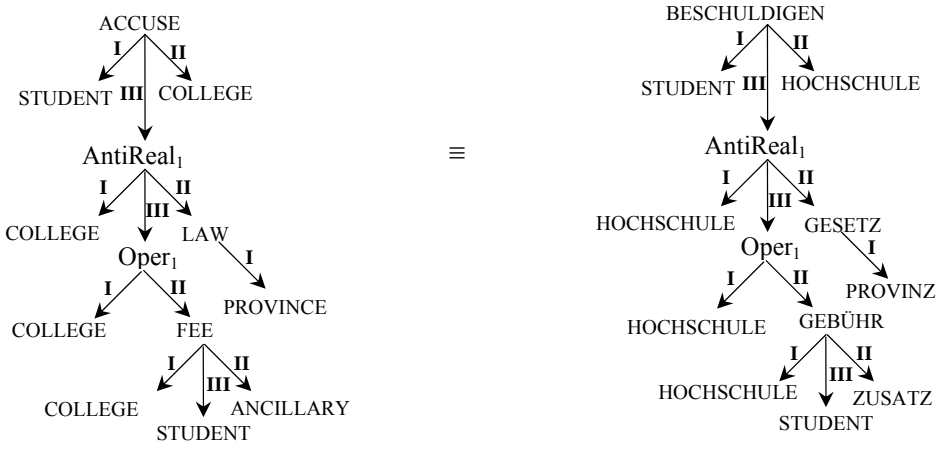


Figure 2: Semantically equivalent English and German DSyntSs

LFs may also serve for resolution of syntactic and lexical ambiguity and for paraphrasing during the transfer procedure (Apresjan et al., 2007).

6 Conclusions

A comprehensive lexical theory that is embedded into a grammatical framework is essential for most of the NLP applications in specialized discourse. We argued that ECL is such a theory. Thus, ECL is comprehensive in that it covers all central lexical phenomena; it is formal enough to be implemented nearly as it is; and it is an integral part of the linguistic model of the MTT. As has been shown in large scale applications (see, e.g., Wanner et al., 2007), MTT's model also supports the incorporation of an extra-linguistic conceptual knowledge representation stratum crucial for specialized discourse, and what is even more important for the terminological nature of specialized discourse. MTT facilitates a rigorous and comprehensive description of conceptual entities, as it does in the case of lexical/terminological entities (Vidal, 2004).

The increasing consideration of MTT in general and of ECL in particular in the field of terminology shows that the potential of the framework has been recognized. It is to be expected that the work of

specialized discourse specialists with MTT will result in new impulses for the theory, making it even broader as it is now.

Acknowledgements: Many thanks to Margarita Alonso Ramos, Marie-Claude L'Homme, Igor Mel'čuk and Vanesa Vidal for many helpful comments on the previous version of this paper. As always, any remaining errors, misinterpretations and other deficiencies are my sole responsibility. The work on this paper has been partially supported by the Spanish Ministry of Science and Innovation under the contract number MCI FFI 2008-06479-CO2-02/FILO.

References

Alonso Ramos, M. (2005) Semantic Description of Collocations in a Lexical Database. In F. Kiefer et al. (Eds.),*Papers in Computational Lexicography COMPLEX 2005*, pp. 17-27. Budapest: Linguistics Institute and Hungarian Academy of Sciences.

Aprejan, Ju.D., I. Boguslavskij, L. Iomdin, A. Lazurskij, V. Sannikov, and L. Tsinman. (1992) ETAP-2 : The Linguistics of a Machine Translation System. *META*, 37(1) :97-112.

Apresjan, Ju.D., I. Boguslavskij, L. Iomdin, and L. Tsinman. (2007) Lexical Functions in Actual NLP-Applications. In L. Wanner (ed.) *Selected Leixcal and Grammatical Issues in the Meaning-Text Theory. In honour of Igor Mel'cuk.* Amsterdam/Philadephia: John Benjamins Publishing Company, pp. 203-233.

Binon, J., S. Verlinde, J. Van Dyck and A. Bertels. (2000) *Dictionnaire d'apprentissage du français des affaires. Dictionnaire de compréhension et de production.*

Bohnet, B. and L. Wanner. (2001) On Using a Parallel Graph Rewriting Grammar Formalism in Generation. In *Proceedings of the 8th European NLG Workshop*, Toulouse.

Bontcheva, K. and Y. Wilks (2004) Automatic Generation from Ontologies: The MIAKT Approach. In *Proceedings of the International Conference on Application of Natural Language to Information Systems*, pp. 324-335.

Budin, G. and Felber, H. (1989) *Terminologie in Theorie und Praxis*. Tübingen: Gunter Narr.

Busemann, S. and H. Horacek. (1997) Generating Air-Quality Reports from Environmental Data. In *Proceedings of the DFKI Workshop on Natural Language Generation*, pp. 15-21.

Cabré, T. (1992) *La Terminologia. La teoria, els mètodes y les aplicacions*. Barcelona: Empúries.

Cabré, M.T. (2003). Theories of terminology. Their description, prescription and explanation. Terminology, 9(2):163-199.

Cabré, M. Teresa (2007). La terminologie, une discipline en évolution : le passé, le présent et quelques éléments prospectifs. In L'Homme, M.-C.; Vandaele, S. (eds.). *Lexicographie et terminologie : compatibilité des modèles et des méthodes*. Ottawa: Les Presses de l'Université d'Ottawa, pp. 79-109.

Claveau, V. and M.-C. L'Homme. (2006) Discovering and Organizing Noun-Verb Collocations in Specialized Corpora using Inductive Logic Programming, *International Journal of Corpus Linguistics*, 11(2):209-243.

Coch, J. (1998) Interactive Generation and Knowledge Administration in MultiMeteo. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pp. 300-303. Niagra-on-the-Lake

Cohen, B. (1986) Méthodes de repérage et de classement des cooccurrents lexicaux. *Traduction et Terminologie*, 2-3 :505-512.

da Cunha, I., L. Wanner and T. Cabré. (2007) Summarization of Specialized Discourse: The Case of Medical Articles in Spanish. *Terminology*, 13(2):249-286.

Daille, B. (2003) Concept Structuring through Term Variations. In *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition, Treatment*. ACL 2003, pp. 9-16. Sapporo, Japan.

Dolbey, Andrew, Michael Ellsworth and Jan Scheffczyk. 2006. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In Bodenreider, Olivier (ed.). Proceedings of KR-MED, pp. 87-94.

Fontenelle, T. (1997) *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen: Niemeyer.

Frawley, W. (1986) Relational Models and Metascience. In M. Evans (ed.) *Relational Models of the Lexicon*. Cambridge: Cambridge University Press, pp. 335-372.

Goldberg, E., N. Driedger and R. Kittredge. (1994) Using Natural Language Processing to Produce Weather Forecasts. *IEEE Expert*, April 1994.

Gough, N. and A. Way (2003) Controlled Generation in Example-Based Machine Translation. In *Proceedings of the Ninth Machine Translation Summit*, pp. 133-140.

Grimes, J.E. (1990) Inverse Lexical Functions. In J. Steele (ed.) *MTT: Linguistics, Lexicography, Applications*. Ottawa: Ottawa University Press, pp. 350-364.

Heid, U. and G. Freibott (1991) Collocations dans une base de données terminologique et lexicale. *Meta* 36(1) :77-91.

Iordanskaja, L., M. Kim, R. Kittredge, B. Lavoie, and A. Polguère. (1992) Generation of Extended Bilingual Statistical Reports. In *Proceedings of COLING '92*, pp. 1019-1022. Nantes.

Janssen, M. (2005) Between Inflection and Derivation. Paradigmatic Lexical Functions in Morphological Databases. In Ju.D. Apresjan and L. Iomdin (eds.) *East West Encounter: Second International Conference on Meaning – Text Theory*. Moscow: Slavic Culture Languages Publishing House, 187-196.

Kageura, K. and M.-C. L'Homme (2008) Reflecting on Fifteen Years of Research and Development in Terminology. *Terminology*, 14(2):153-157.

Kahane, S. and A. Polguère (2001) Formal Foundation of Lexical Functions. In P*roceedings of the Workshop "Collocation: Computational Extraction, Analysis and Exploitation"*, pp. 8-15. Toulouse.

Kittredge, R. and Lehrberger, J. (1982) *Sublanguage: Study of language in restricted semantic domains*. Berlin: Walter de Gruyter.

Kukich, K. (1983) *Knowledge-Based Report Generation: A Knowledge Engineering Approach to Natural Language Generation*. PhD Thesis, University of Pittsburgh.

Laporte, I. and M.-C. L'Homme (1997) Recensement et consignation des combinaisons lexicales spécialisées : exemple d'application dans le domaine de la pharmacologie cardiovasculaire. *Terminologies nouvelles*. 16 :95-101.

Lareau, F. and L. Wanner (2007) Towards a Generic Multilingual Dependency Grammar for Text Generation. In T. Holloway King and E. Bender (eds.) In *Proceedings of the GEAF 2007 Workshop*. Stanford, CA: CSLI Studies in Computational Linguistics On-line (http://csli-publications.stanford.edu).

L'Homme, M.C. (1998) Définition du statut du verbe en langue de spécialité et sa description lexicigraphique. *Cahiers de Lexicologie*. 73(2) :61-84.

L'Homme, M.C. (2007) Using Explanatory and Combinatorial Lexicology to Describe Terms. In L. Wanner (ed.) *Selected Leixcal and Grammatical Issues in the Meaning-Text Theory. In honour of Igor Mel'cuk*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp. 167-202.

L'Homme, M.C. (2008). Ressources lexicales, terminologiques et ontologiques : une analyse comparative dans le domaine de l'informatique, *Revue française de linguistique appliquée*, 13(1), pp. 97-118.

Mel'čuk, I.A. (1988) *Dependency Syntax*. Albany, NY: SUNY Press.

Mel'čuk, I.A. (1996) Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 37-102.

Mel'čuk, I.A. (1997) *Vers une linguistique Sens-Texte* (Leçon inaugurale 139), Paris, Collège de France.

Mel'čuk, I.A. (2006) Explanatory Combinatorial Dictionary. In G. Sica (ed.) *Open Problems in Linguistics*. Monza: Polimetrica Publisher, pp. 225-355.

Mel'čuk, I.A., A. Clas and A. Polguère. (2005) *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Mel'čuk, I.A. and N. Pertsov. (1987). Surface Syntax of English. Amsterdam/Philadelphia: John Benjamins Publishing Company

Mel'čuk, I.A. and A. Polguère. (2007) *Lexique actif du français*. Brussels: De Boeck.

Mel'čuk et al. (1984, 1988, 1992, 1998) *Dictionnaire explicatif et combinatoire du français contemporain*. Vols 1-4. Montréal : Les Presses de l'Université de Montréal.

Mel'čuk, I.A. and L. Wanner (2001) Towards a Lexicographic Approach to Lexical Transfer in Machine Translation. *Machine Translation*. 16:21-87.

Mel'čuk, I.A. and L. Wanner (2006) Syntactic Mismatches in Machine Translation. *Machine Translation*, 20(2):81-138.

Mel'čuk, I.A. and L. Wanner (in print) Morphological Mismatches in Machine Translation. *Machine Translation*.

Mel'čuk, I.A. and A. Zholkovsky. (1984) *Explanatory Combinatorial Dictionary of Modern Russian*. Vienna: Wiener Slawistischer Almanach.

Mille, S., A. Burga, V. Vidal and L. Wanner. (2009) Creating an MTT Tree Bank of Spanish. In Proceedings of the 4th International MTT Conference, Montreal.

Percova, N. (1996). RUSLO: An Automatic System for Derivation in Russian. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 306-317.

Portet, F., E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. (2009) Automatic Gneration of Textual Summaries from Neonatal Intensive Care Data. *Artificial Intelligence*, 173(7-8):889-916,

Rösner, D. (1986) *Ein System zur Generierung von deutschen Texten aus semantischen Repräsentationen*. PhD Thesis, University of Stuttgart.

Sager, N. and Nhàn, N.T. (2002) The Computability of Strings, Transformations and Sublanguage. In B.E. Nevin and S.M. Johnson (eds.) *The Legacy of Zellig Harris*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 79-120.

Shinmori, A., M. Okumura, Y. Morukawa, and M. Iwayama (2003) Patent Processing for Readability. Structure Analysis and Term Explanation. In *Proceedings of the Workshop on Patent Corpus Processing*, pp. 56-65. ACL 2003, Sapporo, Japana.

Temmerman, R. (2000) *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Vidal, V. (2004). Aproximación al fenómeno de la combinatoria verbo-nominal en el discurso especializado en Genoma Humano. Barcelona: Instituto Universitario de Lingüística Aplicada, Universitat Pompeu Fabra.

Vidal, V. (2007) «Consideraciones en torno a la descripción terminográfica de la combinatoria léxica especializada: aspectos macroestructurales ». En M. Lorente, R. Estopà, J. Freixa, J. Martí, C. Tebé (ed.). *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví*. Barcelona: Institut Universitari de Lingüística Aplicada. 473-486

Wanner, L. in print. Report Generation. In N. Indurkhya and F.J. Dalmerau (eds.) *Handbook of Natural Language Processing, Second Edition*. Boca Raton, FL: CRC Press, Taylor and Francis Group.

Wanner, L., B. Bohnet, M. Giereth, and V. Vidal (2005) The First Steps towards the Automatic Compilation of Specialized Collocation Dictionaries. *Terminology*, 13(2):143-180.

Wanner, L. et al. (2007) From Measurement Data to Environmental Information: MARQUIS – A Multimodal Air Quality Information Service for the General Public. In *Proceedings of the 6th International Symposium on Environmental Software Systems*, Prague.

Classes de vocabulaire et indexation automatique : le cas des index de livres

Lyne Da Sylva
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Lyne.Da.Sylva@UMontreal.CA

Résumé

L'indexation automatique peut reposer uniquement sur des méthodes statistiques de calcul de fréquence. Mais des données sur les index préparés par les humains suggèrent qu'à fréquence égale il peut être utile de traiter différemment des mots issus de classes de vocabulaire différentes. Cet article explore cette idée, spécifiquement dans le contexte de la création automatique d'index de livres. Trois classes sont distinguées (vocabulaire courant, vocabulaire scientifique de base, vocabulaire scientifique et technique spécialisé), dont l'utilité diffère dans l'indexation. Une illustration de leur utilisation potentielle est donnée par le biais d'un prototype d'indexation. Les difficultés de constituer ces classes sont exposées. Une expérimentation d'extraction automatique des mots du vocabulaire scientifique de base a permis d'augmenter une liste constituée manuellement.

1 Introduction

Ce travail s'inscrit dans le cadre d'un programme de recherche sur l'indexation automatique, qui porte particulièrement sur l'indexation de type « index de livres »[1]. Pour l'indexation automatique, où l'on fait l'extraction de mots d'un texte afin de faciliter son repérage, on constate que tous les mots n'ont pas un statut égal. À l'examen de quelques exemples d'index produits par des indexeurs humains, on peut voir des exemples comme le suivant :, où trois termes d'indexation sont retenus pour décrire un document :

(1) `Droit communautaire, Application, Jurisprudence`[2]

Avec les termes disciplinaires « droit communautaire » et « jurisprudence », se trouve le mot « application », qui a un sens général (non disciplinaire), soit la nominalisation du verbe « appliquer ». On comprend ici que le document parle de l'application du droit communautaire, et que le sens de « application » ici n'est pas spécifique au droit, mais correspond au sens général du mot. Les indexeurs n'utilisent que parcimonieusement ce type de mot, avec des mots thématiques, pour faire l'indexation des documents.

L'exemple qui suit a la forme d'une entrée d'index que l'on trouve à la fin des livres. Le mot général « application » est utilisé en vedette secondaire, subordonné à une vedette principale spécialisée :

(2) `Droit communautaire,`
`application`

[1] Cette recherché à été effectuée dans le cadre d'une Subvention à la découverte du Conseil national de recherche en sciences naturelles et génie du Canada.
[2] Notice tirée de la base de données RESSAC (Réseau documentaire santé social de l'administration centrale), http://194.199.119.234/Ressac.htm, correspondant à l'article « Dans quelle situation, le droit de l'Union européenne trouve-t-il à s'appliquer en droit interne ? ».

Cette dualité rappelle celle des termes de certains thésaurus documentaires, comme MeSH (*Medical Subject Headings*), qui combinent un descripteur avec un qualificatif. L'objectif du présent article est de défendre la thèse de l'utilité de distinguer certains mots ou certaines classes de mots, en soutien à l'indexation automatique : pour permettre une indexation plus complexe, pour filtrer des éléments lexicaux indésirables et pour privilégier certains types de mots dans certaines circonstances.

La section suivante présente la notion de classes de vocabulaire pertinentes pour l'indexation et la section 3 esquisse une définition théorique pour l'une d'entre elles, le vocabulaire scientifique de base. La section 4 illustre l'utilisation concrète qui peut être faite de ces classes dans un prototype d'indexation automatique alors que la section 5 fait état d'une expérience de dérivation automatique du vocabulaire scientifique de base. La section 6 aborde les problèmes rencontrés dans l'utilisation des classes. Des réflexions et pistes de recherche sont présentées à la section 7.

2 Classes de vocabulaire pertinentes pour l'indexation

Nous préciserons d'abord quelques aspects de la tâche d'indexation ainsi que les travaux de recherche reliés. Puis, nous exposerons le type de classes que nous jugeons utiles, et nous détaillerons comment elles peuvent être utilisées pour l'indexation.

2.1 Types d'indexation et travaux antérieurs pertinents

L'indexation de type « macroscopique » retient un certain nombre de termes utiles globalement pour décrire le document. C'est ce qui est pertinent pour les travaux sur la recherche d'information. Ces termes sont alors normalement en nombre assez limité, entre trois et une vingtaine environ. Il s'agit presque exclusivement de noms, tel que dicté par les normes applicables (ex. AFNOR, 1993). L'exemple (1) en est une illustration. Une bonne partie de la formation en sciences de l'information porte sur ce type d'indexation. On ne s'y intéresse ici que par opposition à l'autre type, l'indexation « microscopique ».

Celle-ci consiste à faire un index tel que l'on trouve à la fin d'un livre (Mulvany, 2005 et Stauber, 2004). Il y a alors typiquement un très grand nombre d'entrées (plusieurs dizaines ou même plusieurs centaines) qui visent à aider l'utilisateur à retrouver un passage précis dans le document. Pour capter tous les sujets abordés ainsi que les différences entre les passages sur des thèmes reliés, les entrées d'index se présentent sous la forme structurée illustrée dans l'exemple (2). Les travaux de recherche sur l'indexation de livres sont très limités; les articles publiés sur le sujet se trouvent presque exclusivement dans des revues professionnelles et représentent surtout des témoignages d'indexeurs. Également, très peu de travaux portent sur l'approche automatique à l'indexation de livres; après Artandi (1963) et Earl (1970), on doit attendre les années 2000 pour voir émerger des travaux parallèles indépendants : Da Sylva (2002, 2004), Da Sylva & Doll (2005) et Nazarenko & Aït El-Mekki (2005).

Dans les deux types d'indexation, une indexation utile du document inclut à la fois des mots simples et des expressions multi-lexicales (« multitermes » dans la tradition d'indexation documentaire), comme « droit communautaire ». Le présent article s'intéresse à la structure des entrées des index de livres, notamment la relation qui unit les vedettes principale et secondaire et le type d'expressions impliquées.

Da Sylva (2004) présente l'éventail des relations sémantiques habituellement présentes dans les entrées structurées des index de livres – voir le tableau 1. Parmi les relations paradigmatiques, la subdivision d'un thème en facettes n'a pas été explorée dans les travaux sur l'indexation de livres (humaine comme automatique). Lesdites facettes mettent en jeu essentiellement des mots comme « développement », « structure », etc. qui partagent certaines caractéristiques sémantiques. L'examen de celles-ci nous a amenés à distinguer les classes de vocabulaire que nous tentons de caractériser. Ces observations trouvent un écho dans certains aspects méthodologiques de l'indexation.

2.2 Définition préliminaire des classes

Pour décrire un document en indexation macroscopique, un indexeur retient des expressions nominales qui répondent à divers critères. En particulier, les termes d'indexation doivent être discriminants, expressifs et précis. Certains mots sont à proscrire, par exemple des mots ambigus et généraux comme

Type de relation	Relation	Exemple
Syntagmatique	Mot – Terme contenant ce mot	Grammaire grammaire de dépendance
Paradigmatique	Hyperonyme – Hyponyme	Mammifères félins
	Tout – partie	Voiture moteur
	Thème – Facette	Robotique développement
	Coordination	Café et grossesse

Tableau 1 : Types de relations sémantiques entre vedette principale et secondaire

« application », « exemple » ou « forme ». Sur la base de ces consignes et suite à l'observation d'index existants (Da Sylva, 2004), on peut diviser le vocabulaire utilisé pour l'indexation en différentes classes :

- le vocabulaire courant (VC) : « appris de façon intuitive, comme toute langue maternelle, surtout composé de termes concrets et de verbes d'action (environ 4 000) » (Waller, 1999, p. 86). Généralement connu des enfants de l'école primaire, il décrit entre autres des objets de la vie courante.
- le vocabulaire scientifique de base (VSB) : « que l'on acquiert dès les études secondaires et qui augmente en importance et en précision au fur et à mesure que l'on approfondit le champ scientifique (4 000 également) » (Waller, 1999, p. 86) - il contient « un très grand nombre de termes dits « généraux » ou « athématiques », tels que : fonction, modèle, opération, système, etc. ». (p. 89). Il sert à exprimer le discours savant.
- le vocabulaire scientifique et technique spécialisé (VSTS) : « propre à chaque discipline, science ou pratique … Il est impossible de les dénombrer, d'autant plus que c'est parmi eux que l'on compte le plus de « naissances » » (Waller, 1999, p. 86). Il fait l'objet des études en terminologie.

La taille de chaque classe est donnée par Waller, qui ne cite cependant aucune source utile pour identifier les listes décrites et en compter le nombre d'éléments. Nous devons nous contenter de cet estimé vague. Notre interprétation de ces classes nous a mené à produire une liste de mots correspondant au VSB. La liste contenait au départ environ 440 mots, identifiés par introspection, analogie, recherche de synonymes de mots déjà identifiés, etc.

Le tableau 2 présente quelques exemples tirés au hasard dans notre liste. Une caractéristique importante : ce sont des mots abstraits (alors que le vocabulaire courant contient essentiellement des concrets). Et ils sont généraux, applicables à tout domaine de la connaissance.

ajustement	avertissement	caractéristique
chose	combinaison	conclusion
contenu	correspondance	corrélation
expansion	figure	forme
hypothèse	infériorité	interprétation
modification	niveau	opération
présence	progression	regroupement
section	source	suite
synthèse	traitement	utilisation

Tableau 2 : Quelques exemples de mots du vocabulaire scientifique de base

Nous recherchons évidemment des critères opérationnels pour distinguer les classes. Les travaux rapportés ici ne sont qu'une première incursion dans cette voie.

2.3 Utilité des classes

L'utilité de chaque classe de vocabulaire varie, comme nous l'avons évoqué ci-dessus. Le VC s'avère souvent peu utile pour l'indexation, car trop polysémique. Au contraire, le VSTS, spécialisé et sémantiquement chargé, contient les meilleurs candidats pour l'indexation. Notons que certains mots normalement classés dans le VC appartiennent, par accident ou par dérive sémantique, à la classe du VSTS pour un domaine donné. Par exemple, les termes désignant les meubles de la maison (comme les tables), que l'on classerait sûrement dans le VC, appartiendraient au VSTS de la menuiserie et seraient utiles pour l'indexation d'un document de ce domaine; nous reviendrons sur ce phénomène de « migration » des mots d'une classe à l'autre pour un domaine donné. Mais retenons que les termes d'indexation discriminants appartiennent en priorité au VSTS ou au VC. Le VSB, quant à lui, est très intéressant en vedette secondaire ou en modification d'entrée, mais peu en vedette principale. Comme sa fréquence d'apparition dans un document n'est pas tributaire du domaine mais du fonctionnement de la langue, une approche statistique est mal adaptée pour intégrer ce type de mots dans l'indexation.

3 Vocabulaire scientifique de base : définition théorique

Parmi les classes présentées, le vocabulaire scientifique de base est particulièrement intéressant, étant donné son utilisation limitée au niveau de sous-vedette.

3.1 Autres travaux similaires

Les travaux sur le vocabulaire sont pour la plupart issus des recherches en acquisition du vocabulaire, en langue maternelle ou seconde. Ainsi, Ogden (1930) a développé un lexique « de base » en anglais (« Ogden's Basic English ») qui comprend des mots issus du VC, du VSB ainsi que des termes que nous classons dans le VSTS. Il ne peut donc servir à nos fins. Les mêmes objections peuvent être faites aux listes de mots extraits par fréquence d'un grand corpus. Elles contiennent de plus des mots de toutes les classes morphosyntaxiques (pronoms, déterminants, etc. inclus). Elles ne sont donc que d'un intérêt limité pour l'indexation automatique.

Les travaux qui se rapprochent le plus des nôtres ont été inspirés de ceux de Phal (1971) sur ce qu'il appelle le « vocabulaire général d'orientation scientifique » (VGOS) :

> « Le vocabulaire scientifique général est (…) commun à toutes les spécialités. Il sert à exprimer les notions élémentaires dont elles ont toutes également besoin (mesure, poids, rapport, vitesse, etc.) et les opérations intellectuelles que suppose toute démarche méthodique de la pensée (hypothèse, mise en relation, déduction et induction, etc.) » (Phal, 1971, p. 9).

Le VGOS de Phal ne contient pas seulement des noms. Il a été défini en étudiant des corpus issus des sciences pures et appliquées et compte 1160 mots. L'auteur note que 63,9% des mots du VGOS appartiennent au français fondamental (Phal, 1971, p. 46), et qu'on y retrouve donc des mots du VC. Drouin (2007) travaille à la définition d'un « lexique scientifique transdisciplinaire », en comparant un corpus de référence de langue générale et un corpus d'analyse en langue spécialisée (scientifique) : d'importantes différences dans les distributions de fréquence suggèreraient des termes « typiquement scientifiques », utiles pour la terminologie. Coxhead (2000) propose le « Academic Word List », une liste de plusieurs centaines de mots organisés en 570 familles. Cette liste vise à soutenir l'apprentissage de l'anglais, niveau avancé. Même si on n'en retenait que les noms, il existe quand même des différences importantes avec nos travaux : comme les 2000 mots les plus fréquents de l'anglais n'y apparaissent pas, la liste ne contient pas « system », « work » ou « study », que nous considérons comme appartenant au VSB de l'anglais.

En bref, pour nos besoins, il est nécessaire d'écarter les mots du VC, de se limiter aux noms et de couvrir non seulement les sciences mais tous les domaines de la connaissance.

3.2 Notre définition

Notre étude nous amène à proposer les critères suivants pour la définition du VSB.

- Caractère général : Les termes du VSB sont applicables dans tous les domaines de la connaissance : sciences pures ou appliquées, sciences humaines et sociales, arts, etc. En effet, les termes « application » ou « début » s'appliquent aussi bien en chimie, en criminologie qu'en cinéma.
- Caractère savant : C'est le vrai « s » dans VSB, puisque les domaines d'application ne sont pas nécessairement « scientifiques ». Nous cherchons donc des termes qui n'appartiennent pas au VC.
- Caractère abstrait : Les termes du VSB dénotent souvent un processus ou un résultat. C'est une observation a posteriori, faite après avoir examiné un grand nombre de candidats. Les mots du VSB peuvent avoir un sens concret, mais ils doivent également avoir un sens abstrait.

Cette définition théorique est nécessaire : elle circonscrit la classe et pourra servira à valider des candidats déterminés de manière automatique. Elle est cependant floue et l'attribution d'un mot au VSB reste relativement subjective. Nous travaillons à l'opérationnaliser.

D'autres critères sont peut-être également pertinents : les mots du VSB tendent à être des prédicats (souvent, des déverbaux ou dé-adjectivaux). Et on peut souvent les reconnaître par leur morphologie : plusieurs se terminent en -ation, -age, -ité, etc., bien que ces deux critères ne soient pas présents pour tous les mots de la classe (« exemple » et « structure », par exemple).

La liste constituée manuellement a bien sûr plusieurs limites, dont la principale est sans doute que nous la savons non exhaustive. Un défi qui se pose est de déterminer une manière de la compléter par des moyens automatiques. Une comparaison entre la liste faite manuellement et une liste créée automatiquement serait sûrement édifiante. Une partie de nos travaux a justement porté sur cet aspect (section 5).

4 Utilisation concrète du VSB dans l'indexation automatique

Malgré les problèmes rencontrés pour la définition de la classe du VSB, la réflexion a permis la construction d'une ressource lexicale utile pour l'indexation automatique.

4.1 Prototype d'indexation automatique

Nous avons développé un prototype d'indexation automatique, Indexo (Da Sylva & Doll, 2005), conçu principalement pour créer des index de livres (en français ou en anglais). Il extrait les noms et multitermes candidats pour l'indexation. L'extraction de terminologie est un sujet de recherche fertile pour lequel plusieurs algorithmes ont été proposés (voir par exemple Bourigault et al., 2001). Nos propres travaux nous ont menés à développer un extracteur de termes compatible avec les visées du prototype. L'extracteur, intégré à Indexo, est basé sur une grammaire d'expressions régulières de catégories morphosyntaxiques. L'indexation repose aussi sur une segmentation automatique du texte.

4.2 Utilisation du VSB pour la création d'entrées d'index de livres

Une première liste de VSB a été établie manuellement, sur la base de la définition sommaire de la section 3.2, afin d'y soutenir l'indexation automatique. Le VSB a été intégré au logiciel d'indexation automatique, pour créer ces entrées d'index complexes faite de la combinaison d'un terme du VSTS et d'un mot du VSB. La méthode utilisée procède ainsi :

- Repérage de mots ou expressions appartenant au VSTS (les détails sont donnés ci-dessous)
- Repérage de mots appartenant au VSB (sur la base de la liste fournie au logiciel)

- Calcul d'un poids selon la distance entre les deux (VSTS et VSB)[3]
- Inclusion dans l'index des paires ayant un poids dépassant un seuil minimum

Cette méthode produit donc un certain nombre de paires « vedette principale – vedette secondaire » proposées comme entrées d'index, comme les exemples 4 à 6 ci-dessous extraits par Indexo d'un texte sur le loup de mer (Desjardins, 2004).

```
(4) loup de mer,
        taille

(5) synthèse annuelle de protéines antigel,
        évaluation

(6) valorisation de la biomasse résiduelle,
        intérêt
```

On voit que les entrées complexes sont plus évocatrices que ne le serait la vedette principale seule.

L'identification du VSTS d'un document est faite sur la base des statistiques de fréquence : les mots les plus fréquents du document constituent a priori la liste du VSTS du document. Certains mots de la liste du VSB peuvent néanmoins se retrouver dans la liste VSTS, selon leur fréquence d'apparition dans le document, s'ils dépassent une fréquence donnée (voir section 4.4)[4].

Une autre application de l'identification du VSB survient lors de l'extraction des termes.

4.3 Utilisation du VSB pour l'extraction de termes

Le processus d'indexation automatique fait l'extraction de mots et multitermes pour cerner le vocabulaire spécialisé. Or, une extraction automatique de termes peut produire des exemples comme en (7) :

```
(7) fabrication de comptoirs de cuisine
    utilisation de béton précontraint
```

Dans ces cas-là, les termes intéressants excluent le premier mot, issu du VSB. Doter le système d'une liste de mots du VSB permet d'en faire l'élagage et ainsi simplifier les termes « comptoirs de cuisine », et « béton précontraint ».

Les deux stratégies peuvent être comparées : on peut évaluer l'expressivité des entres d'index avec et sans VSB. Notre prototype permet de paramétrer l'extraction de termes avec ou sans mots du VSB dans les termes. Il peut donc ou bien extraire « fabrication de comptoirs de cuisine », ou bien extraire « comptoir de cuisine » et le lier au VSB « fabrication » dans la phase de création d'entrées d'index.

4.4 Identification du VSB dans un texte et migration entre les classes

Tel que précisé ci-dessus, le vocabulaire spécialisé est identifié dans Indexo en retenant les mots et expressions les plus fréquents du document. Mais il peut arriver que des mots préalablement inclus dans la liste de VSB appartiennent en réalité au vocabulaire spécialisé d'un texte ou d'un corpus donné.

La liste du VSB sert à identifier des candidats; mais selon les fréquences d'occurrence dans un texte, cette liste doit être révisée contextuellement. Par exemple, « analyse » appartient au VSB. Mais en ma-

[3] On parle ici de distance mesurée en nombre de caractères entre les frontières des termes. La proximité immédiate est privilégiée, comme dans « fabrication de comptoirs de cuisine », mais une certaine distance est tolérée pour capter la relation dans des énoncés plus complexes : « Pour les comptoirs de cuisine, leur fabrication demande une certaine expérience ». Cette tolérance introduit un certain nombre d'erreurs mais aussi une robustesse intéressante.

[4] Nous travaillons pour l'instant sur des documents isolés; l'identification du VSTS d'un document serait grandement améliorée en travaillant sur un corpus thématique.

thématiques, le terme « analyse » décrit une branche des mathématiques qui s'intéresse principalement au calcul différentiel et intégral. Dans un texte ou un corpus en mathématiques, « analyse » appartient plutôt au vocabulaire spécialisé. Il est probable que la fréquence du sens spécialisé y serait anormalement élevée, par rapport aux occurrences du sens général d'« analyse ». À l'aide de l'analyse des fréquences des candidats VSB, on peut en faire l'élagage (et les ajouter aux candidats du vocabulaire spécialisé, ou VSTS) avant de procéder au repérage des paires VSTS-VSB.

Un autre exemple : dans le texte du loup de mer mentionné ci-dessus, le mot « synthèse », à fréquence 6, est au 7e rang (ex æquo avec d'autres) des mots les plus fréquents du document. On y décrit la « synthèse de protéines antigel » du loup de mer. Il ne s'agit pas, pour « synthèse », dans ce contexte, de son sens général « Opération qui procède du simple au composé, de l'élément au tout », ou « Ensemble constitué par les éléments réunis; résultat d'une synthèse », ou encore « Opération intellectuelle par laquelle on rassemble les éléments de connaissance concernant un objet de pensée en un ensemble cohérent; vue d'ensemble qu'on obtient ainsi », qui sont des définitions sans marques d'usage dans le Robert[5]. Il s'apparente plutôt au sens spécialisé en chimie (« Création à partir de plusieurs éléments simples, d'un élément plus complexe »[6]). Il sera plus utile d'identifier le terme complexe « synthèse de protéines antigel » que de fournir une entrée d'index de type VSTS-VSB composée de « protéines antigel, synthèse ». On voit qu'ici l'examen des fréquences donne un bon indice.

En l'absence de désambiguïsation et en l'absence d'un lexique spécialisé pour le domaine, nous avons implémenté l'heuristique suivante dans notre prototype : la fréquence de chaque terme est calculée, et une fréquence seuil est déterminée. Au-delà de ce seuil, un mot est classé dans le VSTS du texte, plutôt que dans le VSB. En outre, puisque le VSTS est identifié sur la base de la fréquence, les termes du vocabulaire courant, s'ils sont fréquents, seront assimilés au VSTS, conformément à leur usage apparent. Une telle utilisation flexible de la liste du VSB s'avère préférable. D'autres mesures pourraient être utilisées pour identifier les « faux VSB » en contexte – voir notamment la section 5.2.

5 Vocabulaire scientifique de base : dérivation automatique

Pour dépasser les limites de l'identification manuelle du VSB, nous avons effectué une expérimentation de dérivation automatique.

5.1 Méthodologie : analyse de corpus

Nous avons procédé par l'exploitation automatique d'un corpus construit pour les besoins de l'étude. Nous supposons que le VSB est présent dans un corpus « suffisamment grand », que les mots du VSB apparaissent indifféremment dans tous les domaines et que les textes utiles pour l'extraire sont accessibles au plus grand nombre de lecteurs possibles mais néanmoins exprimés en langue savante. Nous avons donc constitué un corpus de résumés et titres d'articles savants, tirés de bases de données bibliographiques[7]. Le corpus d'environ 14 millions de mots est en anglais, étant donné les bases de données disponibles (nous n'avons pu reproduire l'expérience pour le français). Nous avons fait l'extraction des noms avec leurs fréquences d'apparition, à l'aide d'un logiciel dédié à la tâche, qui a fait la lemmatisation ainsi que la fusion entre certaines variantes orthographiques. Le comptage s'est limité aux noms et a exclu les mots absents d'au moins une base de données.

---

[5] *Le Nouveau petit Robert de la langue française*, 2007.

[6] *Le Grand dictionnaire terminologique*, http://www.granddictionnaire.com.

[7] ARTbibliographies Modern (arts variés, traditionnels comme innovants), ASFA1 : Biological sciences and living resources (organismes aquatiques, incluant biologie et écologie, exploitation de ressources vivantes, aspects juridiques, politiques et socioéconomiques), Inspec (physique, génie, informatique, technologies de l'information et de la communication, recherche opérationnelle, etc.), LISA (Library and Information Science Abstracts) (bibliothéconomie, sciences de l'information, archivistique et domaines connexes), LLBA (linguistique et disciplines reliées dans les sciences du langage), Sociological Abstracts (sociologie et sciences sociales), Worldwide Political Science Abstracts (sciences politiques et domaines connexes, dont les relations internationales, droit et administration publique).

5.2 Résultats

Par cette méthode, en examinant les 1500 mots les plus fréquents extraits, nous avons réussi à identifier des mots absents de notre liste VSB anglais initiale, mais qui répondent aux critères de définition de la classe. Nous avons donc pu augmenter cette liste initiale de 441 mots à 669 mots (mais nous sommes encore loin des 4 000 postulés par Waller, 1999…).

Une conséquence intéressante de cette approche : comme on a calculé la fréquence moyenne de chaque mot dans le corpus, celle-ci peut servir de fréquence seuil pour départager les mots du VSB des mots du VSTS (problème exposé en 4.2).

La méthode de constitution automatique du VSB possède les limites inhérentes à l'analyse automatique de corpus (faible fréquence, voire absence, de certains mots qui répondent pourtant aux critères définitoires du VSB), mais elle a quand même permis de compléter notre liste existante.

5.3 Autres méthodes

Nous avons exploré la possibilité de constituer le VSB en faisant l'extraction de la différence entre un dictionnaire pour enfants et un dictionnaire pour adolescents. Nous avons cependant rencontré divers obstacles méthodologiques. Notamment, nous n'avons pas réussi à identifier une telle paire de dictionnaires numériques fabriqués par le même éditeur et qui seraient aisément exploitables de manière automatique.

6 Problèmes rencontrés dans l'utilisation du VSB

Deux problèmes importants dans l'utilisation du VSB limitent la performance du prototype d'indexation automatique. D'abord, le VSB n'est représenté que par une liste de graphies, sans sens associé. Pour exploiter réellement le potentiel d'expressivité du VSB, il faudrait un mécanisme de désambiguïsation lexicale en contexte (et savoir reconnaître, par exemple, les occurrences du sens général de « analyse » et celles de ses sens spécialisés). Ensuite, notre méthode permissive relie de façon erronée la paire en (8) à partir de la phrase en (9) :

```
(8) solution de rechange à la pêche,
        développement 1,

(9) Face à cette situation alarmante et à la demande toujours
croissante en produits de la mer, le développement de la mari-
culture, ou élevage d'organismes marins, s'impose comme solution
de rechange à la pêche.
```

Plusieurs raisons expliquent ce choix, liées aux diverses heuristiques à l'œuvre dans le logiciel (et non exposées ici). Mais il est clair que la relation ici est fausse, et qu'il aurait fallu proposer la suivante :

```
(10) mariculture,
        développement 1,
```

La solution évidente serait de restreindre la relation aux termes séparés par la préposition « de » ou de limiter le rattachement au terme le plus proche. Pour l'instant, il est plus utile d'être permissif et d'identifier à la pièce les problèmes afin de proposer une meilleure méthode de repérage des relations.

7 Réflexions et pistes de recherche

La notion du VSB, et plus généralement celle des classes de vocabulaire, méritent davantage de réflexion pour lui permettre véritablement de soutenir les logiciels d'indexation automatique.

7.1 Vers une meilleure définition sémantique du vocabulaire scientifique de base

L'approche de détermination du VSB par statistiques présentée à la section 5 présente un certain nombre de limites. La définition esquissée à la section 3.2 a orienté nos premières recherches, mais elle n'est pas suffisante. Il semble utile de la raffiner par d'autres critères sémantiques, en particulier les contraintes qui semblent opérer sur le ou les actant(s). Un aspect intéressant est la restriction aux actants humains. Par exemple, les mots « réponse », « remarque », « compétence » ou « débat » semblent répondre aux critères de la définition : ils sont généraux, dans la mesure où ils peuvent s'appliquer à tout domaine de la connaissance; cependant c'est un humain qui en est habituellement l'actant principal. Nous avons eu tendance initialement à rejeter les termes qui étaient limités en termes de contraintes sur les actants, mais cela nous semble maintenant un mauvais critère d'exclusion. Au contraire, une voie prometteuse pour une meilleure caractérisation du VSB semble être celle de l'examen de ses caractéristiques sémantiques : contraintes sur les actants (nombre d'actants, types, etc.) d'une part, et sur les traits sémantiques des mots du VSB d'autre part. Cela mènera peut-être à une complexification de la classe.

7.2 Classes de vocabulaire pour l'indexation

Notre travail sur le VSB nous a amenés à réfléchir de manière beaucoup plus globale sur la nature des mots dans les entrées d'index. D'autres classes sont vraisemblablement pertinentes pour l'indexation :

- les mots décrivant le discours : mot, phrase, texte, parole, énoncé
- les temporels : temps, année, décennie, jour, mois, heure, seconde, minute, lendemain, veille, etc.
- les locatifs : périphérie, voisinage, intérieur, dessous, contour
- les termes mathématiques : carré, losange, quotient, somme, ratio, rapport, équation, égalité…
- les unités de mesures : dimensions (longueur, épaisseur, profondeur), statistiques (fréquence, majorité, médiane), unités de mesure (mètre, litre, gramme, pascal, etc.), autres (accélération, altitude, amplitude, durée, pression, vitesse, opacité) …

Ces classes de mots sont peut-être utiles pour la description documentaire, dans le sens où elles peuvent identifier des mots à ne pas inclure dans l'indexation. Mais elles ne sont pas conformes aux critères définitoires du VSB car ces mots possèdent des sens spécialisés.

7.3 Volet applicatif

Nous avons déjà souligné deux applications des classes de vocabulaire dans l'indexation (l'indexation comme telle et l'extraction de termes). Il serait également intéressant d'étudier des propriétés des vedettes utiles pour l'indexation : quelle forme prennent idéalement ces vedettes ? Les rares travaux portant sur cette question (notamment Jones & Paynter, 2001) ne considèrent pas les classes de vocabulaire.

En termes plus généraux, pour la description documentaire, les classes de vocabulaire pourraient aussi servir à l'assistance à la création de thésaurus : pour moduler les résultats d'une opération d'extraction automatique de terminologie et pour alimenter une liste de mots-outils adapté à un domaine de spécialité. On pourrait aussi peut-être structurer les mots du VSB, et constituer ainsi un « thésaurus du VSB ».

Le résumé automatique bénéficierait aussi de l'étude de classes de mots, notamment dans la définition des marqueurs discursifs utilisés pour structurer les textes.

8 Conclusion

Nous avons proposé l'idée d'utiliser de manière différenciée le vocabulaire scientifique de base et le vocabulaire spécialisé d'un document dans la production d'index de livres. Nous avons avancé des méthodes pour constituer ces classes de termes, manuellement et de manière automatique. Nous avons aussi présenté des façons d'utiliser ces classes de termes. Nous avons mené des expérimentations sur ces deux

aspects : une dérivation automatique du VSB (qui nous a permis d'augmenter les termes de notre liste pour cette classe) et l'intégration du VSB à un prototype d'indexation.

Tel que nous l'avons noté dans cet article, plusieurs difficultés se présenteront à qui voudra utiliser les classes de vocabulaire dans une application d'analyse automatique. La polysémie typique des mots appartenant au VSB exigerait une désambiguïsation lexicale en contexte et suscite des questionnements quant à la délimitation opportune des classes. Nous croyons néanmoins qu'il s'agit là d'un terrain inexploré et fertile pour améliorer les performances des algorithmes d'indexation automatique.

Remerciements

Nous désirons remercier les assistants de recherche qui ont contribué à ce travail : Frédéric Doll, Mireille Léger-Rousseau, Nourredine ElmQaddem.

Références

Artandi, S. 1963. *Book indexing by computer*. S.S. Artandi, New Brunswick, N.J.

Association française de normalisation. 1993. *Information et documentation : principes généraux pour l'indexation des documents*. Association française de normalisation, Paris : Afnor.

Bourigault, Didier, Christian Jacquemin, et Marie-Claude L'Homme. 2001. Recent Advances in Computational Terminology, Amsterdam; Philadelphia : John Benjamins.Ogden, Charles Kay. 1930. *Basic English: a general introduction with rules and grammar*. London: Kegan Paul, Trench, Trubner.

Coxhead, Averil. 2000. A New Academic Word List. *TESOL Quarterly*, 34(2):213-238.

Da Sylva, Lyne. 2002. Nouveaux horizons en indexation automatique de monographies. *Documentation et bibliothèques*, 48(4) :155-167.

Da Sylva, Lyne. 2004. Relations sémantiques pour l'indexation automatique. Définition d'objectifs pour la détection automatique. *Document numérique*, Numéro « Fouille de textes et organisation de documents », 8(3) :135-155.

Da Sylva, Lyne et Frédéric Doll. 2005. A Document Browsing Tool: Using Lexical Classes to Convey Information. In : Lapalme, Guy; Kégl, Balász. *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence*, (Proceedings), New York : Springer-Verlag,, pp. 307-318.

Desjardins, Mariève. 2004. *Le loup de mer : une mine d'or moléculaire*. Concours de vulgarisation scientifique de l'ACFAS, http://www.acfas.ca/concours/eureka04/desjardinsmarieve.html

Drouin, Patrick. 2007. Identification automatique du lexique scientifique transdisciplinaire, *Revue française de linguistique appliquée*, 12(2) :45-64.

Earl, L.L. 1970. Experiments in automatic extraction and indexing. *Information Storage and Retrieval*, 6 :313-334.

Jones, S. et G.W. Paynter. 2001. Human Evaluation of Kea, an Automatic Keyphrasing System. In: *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, 148-156.

Mulvany, Nancy. 2005. *Indexing Books*. Chicago : Univ. of Chicago Press.

Nazarenko, Adeline et Touria Aït El Mekki. 2005. Building back-of-the-book indexes. *Terminology*, Special issue on Application-driven Terminology engineering, 11(11) :199-224.

Ogden, Charles Kay. 1930. *Basic English: A General Introduction with Rules and Grammar*. London: Paul Treber & Co., Ltd.

Phal, André. 1971. *Vocabulaire général d'orientation scientifique (V.G.O.S.) – Part du lexique commun dans l'expression scientifique*. Paris : Didier.

Stauber, Do Mi. 2004. *Facing the text : content and structure in book indexing*. Eugene, Or. : Cedar Row.

Waller, Suzanne. 1999. *L'analyse documentaire : une approche méthodologique*. Paris : ADBS.

The equivalence of Internet terms in Spanish and French

Esperanza Valero Doménech
TecnoLeTTra
Universitat Jaume I
Spain
mvalero@trad.uji.es

Verónica Pastor Enríquez
TecnoLeTTra
Universitat Jaume I
Spain
vpastor@trad.uji.es

Amparo Alcina
TecnoLeTTra
Universitat Jaume I
Spain
alcina@trad.uji.es

Abstract

The aim of the DiCoInfo-ES Project is to elaborate a dictionary of computing terms using Explanatory and Combinatorial Lexicology (ECL), taking the French version of the DiCoInfo as a reference. The DiCoInfo is a dictionary of Computing and the Internet that includes the lexico-semantic properties of terms, such as their semantic distinctions and their syntactic and combinatorial functioning. As a first step toward the final objective of the DiCoInfo-Es Project, in this study our intention is to identify the differences between some Spanish and French Internet terms, as far as their form and meaning is concerned. We assume that perfect equivalences between terms in different languages do not always exist. Through this analysis we anticipate the possible problems that could arise when linking the Spanish terms in the DiCoInfo-ES with their French equivalents in the DiCoInfo dictionary.

1 Introduction

In recent years, the number and level of development of specialized fields have grown hand in hand with the progress made in science and technology. However, the number of terminological dictionaries available has not increased proportionally or at the same rate (Alcina, forthcoming).

In addition to traditional dictionaries, translators also use other resources such as huge databases, corpora or translation memories. Nevertheless, we have noted a need for dictionaries which can provide the translator with information quickly and effectively.

The DiCoInfo is a French dictionary, developed by the OLST group from the Université de Montréal, about Computing and the Internet, which takes into account the polysemy of terms, their actantial structures, and semantically related terms. The DiCoInfo is not concept-based, but relies instead on lexical semantics (L'Homme, 2008). This dictionary assumes the principles of ECL, which is the lexicological component of the Meaning-Text Theory. According to the ECL the lexical unit is an entity which has three constituents: a sense, a form and a set of combinatory properties (Mel'čuk et al. 1995: 16). Each sense of a term corresponds to a lexical unit and therefore each sense has a different entry in the DiCoInfo.

In the TecnoLeTTra research group we have started the DiCoInfo-ES Project. In this project we intend to elaborate a Spanish dictionary of Computing and the Internet in which the Spanish terms are linked with the French terms from the DiCoInfo.

In order to design a valid structure which allows us to link the French entries of the dictionary with the Spanish ones, taking into account that there are not always exact equivalences between languages, in this article we compare some Internet terms in Spanish and French, analyzing differences in form, semantic distinctions and the actantial structure of terms.

2 New dictionaries for translators

Among the terminological needs of translators, Cabré (1999: 30) mentions the knowledge of term alternatives and their use in texts, as well as the typical combinatory properties of terms. In addition, Bowker (1998: 648) concludes from a study that corpora are more useful than other traditional resources, such as dictionaries, because terms in corpora are found in the thematic context in which they are normally used. Hence, translators using corpora as terminological resources are less likely to choose a wrong equivalent which cannot be used in the context of the translation.

Moreover, the meaning of a term is determined by the concept it refers to, but also by the context where it is used (Kussmaul, 1995: 89; Robinson, 2003: 113). This is the reason why translators are concerned about looking for terms within their context.

However, dictionaries, even electronic ones, do not have enough contextual information to satisfy the needs of translators. Traditional specialized dictionaries concentrate on answering questions related to the understanding of concepts (L'Homme, 2008) (monolingual dictionaries) or on providing term equivalents without taking into account the context (bilingual dictionaries). Cabré (2004: 105) points out the absence of appropriate dictionaries when searching for an equivalent term in another language. According to this author, the lack of contexts in bilingual dictionaries makes it impossible to choose the most suitable equivalent for each text, especially when the dictionary offers several possibilities.

This lack of contexts in dictionaries is one of the main reasons that some authors (Montero & Faber, 2008: 162) have put forward to explain why some translators, in recent years, have replaced dictionaries with other resources, such as corpora or the Internet. These resources offer large contexts that are easy to access by means of different search techniques that dictionaries have not yet incorporated. Some studies on electronic dictionaries have even integrated corpora into their search systems. A good example is the work by Castagnoli (2008) on the 'EOHS Term' project (http://eohsterm.org). This is a database in which the field *context* is deleted from the entries and terms are linked directly to a reference corpus showing the contexts.

However, if translators have to undertake terminological tasks, for instance searching in a corpus or on the Internet, this will mean wasted time and poorer efficacy in their translations (Alcina, forthcoming). Consequently, there is still a need to create more complete dictionaries which offer quicker access to terminological data and show efficacy in their results. These dictionaries should include, in a systematic way, the contextual information provided by corpora (collocations, arguments, semantic relationships, etc.) and also provide the equivalents in the target language.

3 The equivalence of Internet terms in Spanish and French

In order to meet the challenge of elaborating multilingual dictionaries with contextual information for translators, in the DiCoInfo-ES Project our intention is to design a Spanish dictionary with equivalents connected with the French articles in the DiCoInfo.

The concept of equivalence has been widely discussed in the literature. L'Homme (2004a: 115) argues that terms often represent different realities from one language to another. Rondeau (1984: 32-34) divides the correspondences and differences between terms in different languages into three categories: in the first case, there is an exact correspondence between terms in L_1 and L_2; in the second case, a term in L_1 does not correspond exactly to a term in L_2; and, in the third case, the notion that a term represents in L_1 is not expressed in L_2 because it is an unknown notion or it has not yet received a name. In addition, Dubuc (1992: 55-56) makes a distinction between differences in meaning and differences in use when two terms are not equivalent in two languages.

Differences in equivalence cause difficulties, for example, in the automatic extraction of equivalents from parallel corpora, as shown in the studies conducted by L'Homme (2004a: 208) and Le Serrec (2008: 110), and also in the process of creating multilingual dictionaries. Bae and L'Homme (forthcoming) discuss the adaptations that are necessary in order to convert a monolingual French dictionary into a bilingual French-Korean dictionary with a common methodology designed primarily for French. When assign-

ing the equivalents, three main difficulties arose, namely: differences in the number of constituents in terms, differences in semantic distinctions, and differences in lexical systems. The first difficulty occurs when a single-word term in a language corresponds to a multi-word term or to a phrase in the other language, and vice versa. The second difficulty arises when two equivalent terms in two languages have a different number of senses. For example, a Korean term has two different senses and each sense corresponds to a different French equivalent. The third difficulty is caused by the differences between the lexical systems, for example, syntactical or morphological differences. In order to overcome these differences, they designed an intermediate table which allowed users to observe exact equivalents and at the same time access articles in each database.

Fig. 1. Intermediate table (Bae & L'Homme, forthcoming)

Korean	I-Korean	I-French	French
클릭 1 'click'	클릭할 수 있는 'possible to click'	cliquable 1	cliquable 1 'clickable'

For the inclusion of a Spanish module in the DiCoInFo, as with Korean, a number of problems could arise. These may affect the structure of the dictionary, that is, in the way equivalent articles in Spanish and French will be linked:

- The absence of a term in L_2 which expresses the same meaning as a term in L_1. For example, in French there is a term which describes a set of programs: Mult (logiciel) = logithèque, whereas in Spanish there is no equivalent term for *logithèque*.
- A term in L_1 corresponds to a phrase that cannot be considered a term in L_2. As L'Homme (2008) states, the French term *cliquer* (to click) corresponds preferentially to the Spanish expression *hacer clic* (literally to do click).
- A term in L_1 expresses more senses than its equivalent term in L_2.

Thus, the aim of this article is to perform a comparative analysis of Spanish and French Internet terms, taking into account ECL. We will compare semantic distinctions, differences in form, and the actantial structures of Spanish and French terms. The actantial structure describes the essential participants which describe the sense of a term. For example, the actantial structure of the verb *to configure* has two actants: somebody ~ something (L'Homme, 2008). The semantic distinctions refer to the senses that a term can have in a specialized domain. For example, as L'Homme (2004b) states, the term *program* can have three different meanings: "1) a set of instructions written by a programmer in a given programming language in order to solve a problem (this meaning is also conveyed by computer program); 2) a set of programs (in sense 1) a user installs and runs on his computer to perform a number of tasks (this meaning being also conveyed by software program); and 3) a small set of instructions designed to run a specific piece of hardware".

This is a pilot study to anticipate problems that we will come across when building the Spanish version of the DiCoInfo.

3.1 Methodology

The methodology of this analysis can be divided into five stages: selection of French terms, search for Spanish equivalents, selection of Spanish contexts, linguistic description of Spanish terms, and comparison of Spanish terms with the French version of the DiCoInfo. These different stages are detailed below.

Selection of French terms: We chose the subfield 'Internet' within the field of 'Computing' in order to limit the number of terms to be analyzed. We analyzed terms related to entities and activities that occur on the Internet, such as *anti-logiciel espion* (antispyware), *portail web* (Web portal) or *fureter* (to surf), but we discarded other terms closer to the software or hardware areas, such as *barre espace* (space bar) or *décompacter* (to unzip).

We extracted the list of terms for this subfield by comparing the terms included in the DiCoInfo and the terms included in the *Vocabulaire d'Internet*[2] of the Banque de terminologie du Québec elaborated by the Office québécois de la langue française. In doing so, we obtained all the terms from the DiCoInfo that were also included in this specific glossary about the Internet. The result was a list of 231 terms in French. Some synonyms and morphological variations that we found in the DicoInfo during the analysis were added to this list, which finally gave us total of 242 terms.

Search for Spanish equivalents: We obtained the equivalent Spanish terms by searching in different bilingual dictionaries and glossaries, and also by using our intuition and experience as Internet users. The equivalents found were confirmed by searching in Spanish corpora about Computing.

Selection of Spanish contexts: We obtained contexts for these equivalents in three different corpora.

- A Computing corpus in Spanish, elaborated by the OLST group from the Université de Montréal and containing about 500 000 words.
- Online Corpus Bwananet, elaborated by IULA from the Universitat Pompeu Fabra (Spain), containing more than one million words about computing.
- Corpus de Referencia del Español Actual, an online corpus elaborated by the Real Academia de la Lengua Española, in which we selected the area of Computer science.

We would like to clarify that, since our objective is to create a dictionary for translators, we have only considered corpora which include mostly normative language, whereas the French terms of the DiCoInfo were extracted from a varied corpus. This implies some constraints regarding the selection of Spanish terms to be included in the DiCoInfo, and it has an influence on this analysis as well.

Linguistic description of Spanish terms: In this stage we described the Spanish terms that we found, in terms of their formal and semantic aspects and their linguistic use in contexts.

- Description of the term form: On the one hand, we indicate whether it is a single-word term and, if this is the case, whether it contains a prefix or a suffix. On the other hand, we also indicate whether the term is made up of a number of words (multi-word term) and, if so, we describe the morphological characteristics of its components. For instance, the Spanish term *programa gratuito* (freeware) consists of a noun plus an adjective. There are some cases in which a French term does not have a direct equivalent term in Spanish and instead a phrase is used which cannot be considered a term. For example, the French term *spammer* (to spam) could be in Spanish *enviar spam* (to send spam).
- Description of semantic distinctions: By looking at the Spanish concordances of terms we established whether a lexical form in French corresponds to one lexical unit, whereas its Spanish equivalent corresponds to two or more lexical units. For instance, in Spanish *diálogo* can refer to the dialog between two programs or computers, to the dialog between a user and a computer or to the box that appears on a display screen.
- Actantial structure description: We determined the actantial structure for each lexical unit. For example, the term *diálogo* needs two actants: an agent (user) and an addressee (computer, program).

Comparison of Spanish terms with the French version of the DiCoInfo: Once we had collected the linguistic data about the Spanish terms, we compared them with the information provided by the Di-

[2] This dictionary can be accessed at
http://www.oqlf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/Internet/Index/index.html

CoInfo about the French terms. In this stage we looked for differences relating to semantic distinctions, number of synonyms, formal structure and number of actants of a lexical unit.

3.2 Results

This section contains the results of our comparative analysis of Spanish and French terms. Before going on, we would like to clarify that this analysis was carried out from the information found in the DiCoInfo of French terms and the concordances of Spanish terms from the corpora that we mentioned in this article. In this regard, meanings and uses other than the ones we show here may also be possible, both in Spanish and French.

Of the 242 terms which were analyzed, 201 were nouns, 36 were verbs and 5 were adjectives.

There are many cases in which we did not find interlinguistic variation and no differences were found in the actantial structure of Spanish and French terms. For example, *binette* in French and *emoticón* in Spanish have the following actantial structure: an agent, i.e. a user, and a destination, i.e. a message. As regards semantic distinctions, we did not observe any differences in verbs and adjectives, and only a few differences in nouns. In the rest of the phenomena analyzed there were no differences among the five adjectives.

In the following sections, we will focus on the interlinguistic variations found in the analysis. We only include some examples that illustrate these differences: the rest of the data can be found in Appendix A.

A. **Differences in semantic distinctions:** Most of the French nouns have parallel senses in Spanish, i.e. *accès à Internet* and *acceso a Internet* (Internet access), *adresse de courriel* and *dirección de correo* (e-mail address), *nom d'utilisateur* and *nombre de usuario* (user name), etc. However, we did find 5 French terms with different semantic distinctions in Spanish, that is, one language has more semantic distinctions than the other. For instance, *annuaire* in French is a directory of websites, whereas its Spanish equivalent *directorio* is a directory of websites, but also a group of files. Another example is the French term *journal* which only has the meaning of log, whereas its Spanish equivalent *registro* means a log, the action of introducing data, a database entry, and the process of creating an account on a website (to sign up).

B. Use of synonyms: The number of synonyms and quasi-synonyms of a term also varies from one language to the other. In French we found many cases of synonymy. For example, whereas in Spanish there is only the term *chat* to designate a virtual room where a chat session takes place, in French there are many synonyms: *bavardoir, bavardoir électronique, clavardoir, espace de bavardage, salon de bavardage, salon de clavardage* or *salon de cyberbavardage*.

As in nouns, synonymy is more frequent in the French verbs: *chater, clavarder, cyberbavarder, bavarder* (to chat); whereas in Spanish only *chatear* (to chat) or *conversar en un chat* (to talk in a chat room) are used.

C. Use of phrases which cannot be considered terms: We found some French terms which are equivalent to a phrase in Spanish. For instance, the French noun *blogage* (blogging) can be translated in Spanish as *participación en un blog* (taking part in a blog), *interacción en un blog* (interacting in a blog), etc.; *bloggeur* (blogger) can be translated as *usuario que dispone de un blog* (user of a blog) or a similar expression. Other examples are *polluposteur* and *spammeur* (spammer), in Spanish *emisor de spam* (spam sender), or *spammage* and *pollupostage* (spamming), in Spanish *envío de spam, envío de correo basura* (sending of spam).

As regards verbs, we found six verbs in French that are single-word terms while in Spanish they correspond to phrases which cannot be considered terms:

- *Sauvegarder* (to back up) corresponds to *hacer una copia de seguridad* (to make a backup).
- *Bloguer* (to blog) does not have an equivalent term in Spanish. We use some expressions such as *participar en un blog*, *interactuar en un blog*, (to participate or to interact in a blog).
- *Polluposter* or *spammer* (to spam) correspond in Spanish to the expression *enviar spam* (to send spam).

D. Differences in form: We classified the differences in form into morphological variants and differences in the formal structure.

- Morphological variants: According to the analysis in French there are many morphological variants of the same term, whereas this is not the case in Spanish. For instance, in order to designate the creator or user of a blog in French you can use *webblogueur, webloger, weblogger, webloggueur, weblogueur, bloger, blogeur, blogger, bloggeur, bloggueur*.

- Formal structure of nouns: Some nouns are made up of a single component in one language whereas in the other language their equivalents are multi-word terms. For example, *sauvegarde* (backup) is a single-word term equivalent to the multi-word term *copia de seguridad* in Spanish.

4 Discussion

From this analysis we identified many similarities between the Spanish and French terms in the field of the Internet, such as the actantial structure, which shows no variation in the terms that were analyzed. However, this analysis also shows many differences between the terms in the two languages. Some of these differences can be due to the nature of the corpora used in the analysis. The language of the Spanish corpora seems to be more restricted than the French. For example, we found numerous morphological variants in French (i.e. *webloger, weblogger, webloggueur*), which shows that the texts included in the French corpus are more diverse. In addition, the high frequency of use of phrases which cannot be considered a term in Spanish may be due to the fact that the Spanish corpora include mostly the Spanish of Spain and formal language. For instance, the term *bloguear* (to blog) is not found in the corpora, although it is frequently used in some forums on the Internet. This means we will have to be very careful when deciding on the criteria we will take into account when compiling the Spanish corpus for the DiCoInfo.
Yet the following differences are worth taking into account when elaborating the Spanish version of the DiCoInfo:

- a French term has only one sense and its equivalent in Spanish has more senses or vice versa, i.e. in the DiCoInfo only one meaning was considered relevant for the term *journal* (log) in French, whereas in Spanish the equivalent *registro* has different meanings in this domain.
- in French, the DiCoInfo includes a wide range of synonyms for some terms that in Spanish have no synonyms or only a few, i.e. *clavarder*, *bavarder*, *cyberbavarder*, *chater* (to chat) correspond to *chatear* in Spanish.
- a French term corresponds to a phrase which cannot be considered a term in Spanish, i.e. the French verb *spammer* (to spam) in Spanish would be *enviar spam* (to send spam).
- a single-word term in French has a multi-word equivalent in Spanish, i.e. *sauvegarde* (backup) is equivalent to *copia de seguridad*.

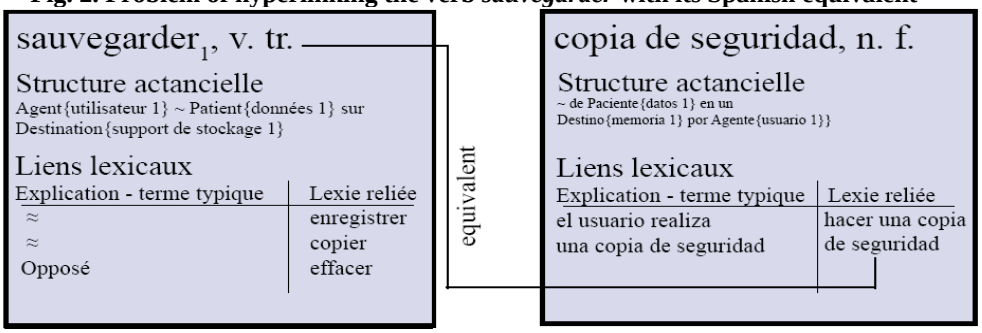
These differences will entail three main problems when building the Spanish version of the DiCoInfo: the extraction of the candidate terms, the selection of headwords, and the hyperlinking between the Spanish and the French entries in the DiCoInfo.

Regarding the extraction, some phrases which cannot be considered a term in Spanish are composed of a general verb and a specific noun. When applying the DiCoInfo methodology, in which only single-word terms are extracted, the general component of the phrases will not be recorded high on the list of candidates or it will not even be extracted. For example, the equivalent of the French *sauvegarder* (to back up) in Spanish is *hacer una copia de seguridad*. The problem is that the verb *hacer* in Spanish has a general meaning and it will not be proposed as a candidate term in the specialized fields.

As far as selection of headwords is concerned *sauvegarder* is a headword in the DiCoInfo whereas the equivalent phrases will not be headwords in the Spanish version.

As regards hyperlinking, the verb *sauvegarder* (to back up) would be linked to the noun *copia de seguridad* (backup) instead of its equivalent verb.

Fig. 2. Problem of hyperlinking the verb *sauvegarder* with its Spanish equivalent



This analysis shows that there are no difficulties that cannot be overcome when building the Spanish dictionary, and indeed there are many similarities between French and Spanish terms which will allow the methodology of the DiCoInfo to be applied in elaborating the Spanish version. However, the problems and the linguistic phenomena described above will require some adaptations to the Spanish module of the DiCoInfo before it can be connected with the French module.

5 Conclusion

The comparison of equivalent terms from the field of the Internet in French and Spanish shows many similarities, but also a wide range of differences relating to semantic distinctions, differences in form and equivalence of French terms with Spanish phrases which cannot be considered terms. Therefore, when connecting the Spanish module of the DiCoInfo with the French module, it is not enough just to link a Spanish term with its equivalent term in French. This will require the design of a complex system capable of representing and relating the links between the Spanish and the French articles.

For this reason, some future work is planned. As well as reflecting on the criteria to be used in compiling the Spanish corpus and writing the articles of the Spanish terms following the methodology of the DiCoInfo, we also plan to develop strategies in order to overcome the difficulties that have arisen from this study. To do so, we will take into account the solutions adopted in the creation of the Korean module of the DiCoInfo (Bae & L'Homme, 2008), such as the design of an intermediate equivalence structure to act as a bridge between the Korean and the French entries. We will reflect on how to apply this strategy between Spanish and French, as well as other possible solutions.

References

Alcina Caudet, Amparo (forthcoming). Metodología y tecnologías para la elaboración de diccionarios terminológicos onomasiológicos. In Amparo Alcina Caudet et al. (Eds.), *Terminología y sociedad del conocimiento*. Bern: Peter Lang.

Bae, Hee Sook & Marie-Claude L'Homme (2008). Converting a Monolingual Lexical Database into a Multilingual Specialized Dictionary. In Frank Borres et al (Eds.) *Multilingualism and Applied Comparative Linguistics* (pp. 213-230). Newcastle: Cambridge Scholars Publishing.

Bowker, Lynne (1998). Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. *Meta* 43(4), 631-651.

Cabré, Mª Teresa (1999). Fuentes de información terminológica para el traductor. Técnicas documentales aplicadas a la traducción. In María Pinto & José Antonio Cordón (Eds.), *Técnicas documentales aplicadas a la traducción* (pp. 19-39). Madrid: Editorial Síntesis.

Cabré, Mª Teresa (2004). La terminología en la traducción especializada. In Consuelo Gonzalo García & Valentín García Yebra (Eds.), *Manual de documentación y terminología para la traducción especializada* (pp. 19-39). Madrid: Arcos/Libros.

Castagnoli, Sara (2008). Corpus et bases de données terminologiques: l'interpretation au service des usagers. In François Maniez et al. (Eds.), *Corpus et dictionnaires de langues de spécialité* (pp. 213-230). Bresson: Presses Universitaires de Grenoble.

Dubuc, Robert (2002). *Manuel pratique de terminologie*. Montreal: Linguatech.

Kussmaul, Paul (1995). *Training the Translator*. Amsterdam/Philadelphia: John Benjamins.
L'Homme, Marie-Claude (2004a). *La terminologie: principes et techniques*. Montreal: Les Presses de l'Université de Montréal.
L'Homme, Marie-Claude (2004b). *A Lexico-semantic Approach to the Structuring of Terminology.* Paper presented at the CompuTerm 2004: 3rd International Workshop on Computational Terminology, Geneva.
L'Homme, Marie-Claude (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire* 217, 78-103.
Le Serrech, Annaïch (2008). *Étude sur l'équivalence de termes extraits automatiquement d'un corpus parallèle : contribution à l'extraction terminologique bilingue*. University of Montreal master's thesis.
Mel'čuk, Igor A., André Clas & Alain Polguère (1995). *Introduction à la lexicologie explicative et combinatoire.* Brussels: Duculot.
Montero Martínez, Silvia & Pamela Faber Benítez (2008). *Terminología para traductores e intérpretes*. Granada: Tragacanto.
Robinson, Douglas (2003). *An Introduction to the Theory and Practice of Translation*. London: Routledge.
Rondeau, Guy (1984). *Introduction à la terminologie*. Chicoutimi: Gaëtan Morin.

Acknowledgements

We would like to thank Marie-Claude L'Homme for her comments and suggestions on this article.

This research is part of the Ontodic and DicoInfo-ES projects, funded by Ministry of Education and Science (TSI2006-01911) and Fundación Universitat Jaume I-Bancaixa (P1-1B2008-57).

Appendix A. Differences of the analysis

Table 1. Differences in the semantic distinctions of nouns

French term	**French semantic distinctions**	**Spanish term**	**Spanish semantic distinctions**
annuaire	1. website directory	directorio	1. website directory 2. an organizational unit, or container, used to organize folders and files
courriel	1.an electronic message 2. group of e-mails (mass noun) 3. e-mail technology	e-mail correo	1. an electronic message 2. group of e-mails (mass noun) 3. e-mail technology. 4. e-mail address.
dialogue	1. dialog between two programs or computers 2. dialog between a user and a computer	diálogo	1. dialog between two programs or computers 2. dialog between a user and a computer 3. a box that appears on a display screen
journal fichier journal	1. a log	registro	1. a log 2. action of introducing data 3. database entry 4. the process of creating an account on a website (sign up)
port	1. physical port	puerto	1. an interface on a computer to which you can connect a device 2. an endpoint to a logical connection

Table 2. Use of synonyms in nouns

French synonyms	**Spanish synonyms**
a commercial arobas/arrobas	arroba
antiespiogiciel antispyware anti-logiciel espion	antispyware

French synonyms	Spanish synonyms
bavardoir bavadoir électronique clavardoir espace de bavardage salon de bavardage salon de clavardage salon de cyberbavardage	chat
binette émoticône frimousse	emoticón emoticono
bloc-notes weblog blog/blogue joueb carnet web	blog bitácora diario web cuaderno de bitácora
weblogueur/webloger/weblogger webloggueur/weblogueur bloger/blogeur/blogger/bloggeur//bloggueur	usuario de un blog
pare-feu coupe-feu	cortafuegos
furetage surf navigation	navegación
graticiel/gratuiciel freeware logiciel gratuit logiciel freeware	software libre freeware software de libre distribución
pollupostage spammage spamming	envío de spam
spam pourriel polluriel	spam correo basura

Table 3. Use of synonyms in verbs

French synonyms	Spanish synonyms
chatter/chater clavarder cyberbavarder bavarder	chatear conversar en un chat
polluposter spammer	enviar un spam
télécharger	descargar bajar

Table 4. Use of noun phrases

French term	Spanish phrases
blogage	conversación en un blog
blogueur	persona que utiliza un blog
clavardage bavardage cyberbavardage	conversación en un chat
graticiel gratuiciel	programa de libre distribución software libre programa gratuito
polluposteur spammeur	generador de spam emisor de spam
pollupostage spammage	envío de spam
webmestre	administrador del sitio

Table 5. Use of verb phrases

French term	Spanish verb phrases
sauvergarder	hacer una copia de seguridad
clavarder bavarder	conversar en un chat hablar en un chat or internet
polluposter spammer	enviar un spam
bloguer	interactuar en un blog participar en un blog

Table 6. Differences in form for nouns. Morphological variations

French terms	Spanish terms
arobas arrobas	arroba
weblog blog blogue	blog
weblogueur webloger weblogger webloggueur weblogueur bloger blogeur blogger bloggeur bloggueur	usuario de un blog
graticiel gratuiciel	software libre

Table 7. Differences in form for verbs. Morphological variations

French terms	Spanish terms
chatter chater	chatear

Table 8. Differences in form for nouns. Number of constituents

Spanish single-word term	French multi-word term
cortafuegos	coupe-feu
adjunto	pièce jointe
contraseña	mot de passe
encabezado	en-tête

French single-word term	Spanish multi-word term
pourriel, polluriel	correo basura
sauvegarde	copia de seguridad
webmestre	administrador web, administrador del sitio

The Framing of Surgical Procedure Terms

Maria-Cornelia Wermuth
Lessius/Katholieke Universiteit Leuven
Department of Applied Linguistcs
Sint Andriesstraat 2
B-2000 Antwerp

cornelia.wermuth@lessius.eu

Abstract

Medical classification systems are used for the conversion of medical natural language data into standardized numeric codes in order to allow for their electronic processing. Our paper focuses on the descriptions of codes in natural language. Code descriptions are produced from and for medical professionals and have specific syntactic and lexical features. However, in daily practice poor coding results are frequent. One of the main reasons is the fact that medical coding is predominately performed by professional coders lacking the medical background knowledge necessary for interpreting code descriptions properly. In order to tackle this problem we propose a frame-based representation model of code descriptions which is supposed to support the coding practice in different ways. We illustrate by means of corpus samples how the computer implementation of the proposed model efficiently supports the explicitation, translation, terminological standardization and generation of code descriptions in different languages.

1 Introduction

This paper deals with a frame-based and cognitively inspired approach to code descriptions or so-called classification rubrics describing surgical procedures. It is based on earlier research into the conceptual structure of classification rubrics (Wermuth 2005) in which among others the advantages of frames in terms of semantic disambiguation and term translation have been investigated. For the present exploration we adopted a usage-based approach (Gries & Stefanowitsch, 2006). We analyzed the conceptual structure of a parallel translation corpus consisting of rubrics describing cardiovascular procedures in English (source language), Dutch, French and German (target languages). The corpus has been taken from the procedure volume of the ICD-9-CM (International Classification of Diseases, 9th revision, Clinical Modification), which is used in hospitals for the registration of surgical procedures. Before we describe the results of this analysis and our frame-based approach to rubrics in greater detail, we first sketch some characteristic features of medical classifications and rubrics and the challenges of rubrics with respect to the coder.

1.1 Medical classifications and rubrics

From a terminological viewpoint medical classifications can be defined as knowledge representation systems in which symbols (such as codes, terms, abbreviations, etc.) refer to objects and states of affairs in the real world. These objects and states (being body parts, medical instruments or operations) are hierarchically classified into subclasses according to criteria dependent on the purpose of the classification. The procedure volume of the ICD-9-CM, which we used for the present investigation, is designed for statistical purposes. As a consequence, the objects (being procedures) are grouped into statistical relevant classes according to specific characteristics such as anatomical criteria, the kind of

procedure, its frequency, etc. The relationship between codes and rubrics and the objects they refer to occurs through concepts or mental representations within the mind. The adopted version of the well-known semantic triangle (Figure 1) reflects these indirect relationships.

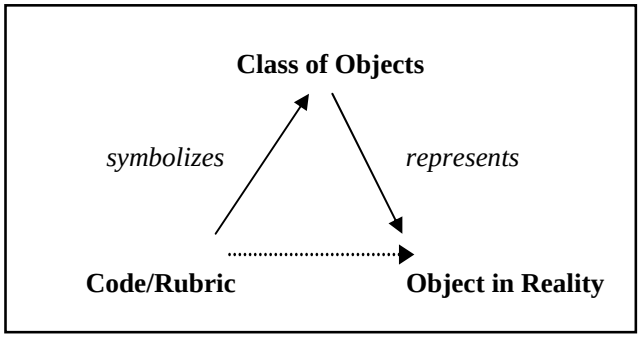


Figure 1. The Relationships between Objects, Codes and Classes of Objects according to the Semantic Triangle.

1.2 What is coding?

Coding means that a human coder is linking data contained in natural language documents such as surgical reports with corresponding classification codes. This activity implies that individual objects (or so-called instantiations of an object class) such as the replacement of the aortic valve, which is described in the surgical report, are assigned to a predefined class of objects on the basis of similarity. In this process first the linguistic data in the source document are to be interpreted and, in a second step, the data in the corresponding classification rubric. In practice, this is a nontrivial task, as we will demonstrate in the following.

Surgical reports as well as rubrics represent specialized texts which refer to the same state-of-affairs or in Langacker's (1987: 63ff.) terms to the same 'scenarios'[1] in the outside world, namely complex surgical procedures. However, this reference is realized in another way due to the different communicative function of both text types. Written operative reports serve the accurate documentation of procedures and should therefore be as detailed as possible (including, for example, information on the involvement of other organs and tissues, the healing or palliation of symptoms, the location and extension of tumors, etc.). Rubrics, by contrast, have a restricted pragmatic function which is to predicate as efficiently as possible the meaning of numerical codes in order to support coding. Rubric terms serve so to say as cues for the coder to access the appropriate procedure scenario and to reconstruct the meaning of the code. From the referred procedure scenario only those aspects of information are selected and linguistically realized which are relevant in the context of the classification system. For example, the operative approach, the operated body parts or the frequency of the operation are aspects of a procedure which are relevant for statistical oriented classifications such as the ICD. By contrast, the subject specifying the initiator agent (i.e. the surgeon taking the initiative and control of the action) is not mentioned as this knowledge is an inherent part of the conceptual representation of the operation scenes evoked by rubrics and, in addition, irrelevant in classification terms. In other words, coding always involves a loss of information as rubrics are the result of a selective purpose-dependent process. It may be obvious that the

[1] These 'scenarios' can be assumed to be 'universal' given the heavily conventionalized way in which medical objects and actions (e.g. surgical procedures in the given case) are conceptualized independent from any socio-cultural contexts.

expert-to-expert discourse triggered of by rubrics only can be successful if there is a common ground of shared (specialized) knowledge between the designer and the addressee (the coder).

1.3 The morphosyntactic features of rubrics

The implicitness of rubrics is reflected by their surface structures. In morphosyntactic terms, rubrics are across all the languages under investigation reduced phrasal forms of (complex) sentences resulting in nominal phrases and a sequence of prepositional phrases. In their most simple realization, rubrics consist of a nominalized action verb (e.g. transplantation, replacement, insertion) or a neoclassicial root form (e.g. -tomy, -plasty, -ectomy) which is complemented by a number of premodifications such as the examples given in (1), (2) and (3).

(1) Heart transplantation
(2) Closed heart valvotomy
(3) Single vessel percutaneous transluminal coronary angioplasty

The head may also be followed by an of-phrase which introduces the direct object (DO) like in (4), and the DO, in turn, may be followed by other prepositional phrases (PP) like in (5):

(4) Replacement of heart valve
(5) Replacement of mitral valve with tissue graft

Apart from the fact that rubrics frequently are quite complex nominal phrases (cf. the example given in 6), in each language under scrutiny they also display a number of other typical features such as the omission of parts of nominal constituents (ellipses) like in (7), the use of paralinguistic markers such as elliptical appositions (sometimes within parentheses) like in (8), the use of polysemous prepositions like in (9), the use of conjunctions or disjunctions in combination with omission like in (10), and the frequent use of unmotivated lexical and orthographical variants like in (11).

(6) Open chest coronary artery angioplasty
(7) Replacement of aortic valve with tissue graft
(8) Insertion of heart assist system, not otherwise specified (NOS)
(9) Excision of aneurysm of heart
(10) Studies of registration and electrophysiological stimulation of heart
(11) Septal defect of heart vs. Septal defect in heart

This elliptical nominal structure of rubrics is derived from syntactically deeper structures of a different form. This form is more closely reflected by the complex surface sentences in the detailed procedure descriptions in surgical reports. Here, the active verbal description highlights the temporal sequence of the actions. In rubrics, by contrast, we see that the basically sequential mode of the referent scenes is presented from a synoptic perspectival mode. Instead of active verbs, action nominalizations or bound neoclassical morphemes (cf. examples given in 12-16) are used to express the dynamic actions which operations basically are. In the verbal construction the act is represented in terms of ‘Agent affecting Patient’, whereas its nominalization makes that the referent is reified and conceptualized as an object. The effect of the absence of an active verb is a sense of a depersonalized and static event.

(12) Heart revascularization by arterial implant
(13) Replacement of heart valve
(14) Closed heart valvotomy
(15) Cardiotomy
(16) Pericardiectomy

2 The challenges of rubrics

Investigations reveal that in practice medical data quite often are not correctly coded. This has a number of causes such as the huge size of most classifications, the high complexity of rules to be followed in order to code correctly, structural shortcomings of classifications or the fact that coders always are very pressed for time. Besides these causes we can assume that also the surface structure of rubrics has an import share in the misinterpretation of codes. As described above (cf. subsection 1.2) the main function of rubrics is to convey information about the code's meaning in an efficient and economic way. Medical experts design rubrics for other experts whose knowledge of and around the theme is supposed to be the same. This explains the telegram style which, in fact, is effective in the intended setting in which the medically trained recipient of the 'rubric message' is able to fill in the gaps on the basis of his or hers domain knowledge and to reconstruct the conceptual structures of the designated objects and state of affaires. The problem, however, is that in medical practice the coding task is mostly not accomplished by domain specialists, but by professional coders (i.e. non-specialists such as staff members without medical training). Having a limited access to the overall conceptual context in which the information encoded in rubrics has to be integrated in order to be correctly interpreted, they more often than not make incomplete or even incorrect analyses of the intended meaning of rubrics with wrong code assignments as a result. The example in (17) may illustrate this.

(17) Replacement of any type pacemaker device with dual chamber device

The interpretation of the PP 'with dual chamber device' in this rubric has two plausible readings of which only the second one is correct in the given medical context. The first reading starts from the assumption that the PP is an adjunction of the verbal phrase 'replacement' which results in the instrumental reading that any pacemaker (i.e. not necessarily a two chamber pacemaker), which has been implanted at an earlier stage is replaced by means of a pacemaker device that supports the stimulation of the two heart chambers. In the second reading, which is the correct one, the PP is interpreted as a modifier of the NP 'pacemaker device'. The rubric's meaning then is that all types of two heart chamber devices are replaced. In this reading, the surgical devices used to perform this operation are simply gapped. The example in (18) illustrates another tricky aspect of rubrics, namely the fact that the predicate-argument-structure is not necessarily identical to the procedure which is actually referred to.

(18) Revision of corrective procedure on heart

Without sufficient background knowledge, the predicate-argument-structure of this rubric suggests that 'revision' is an operative action, which is carried out on a 'corrective procedure on the heart', whatever this could mean. In (medical) reality, the term 'revision' refers to the purpose of a whole procedure which is conducted on the heart or some part of it in order to control the effectiveness of a previously performed procedure.

From the above it may be clear that rubric terms are associated with a conglomerate of terminological, domain-specific and encyclopedic knowledge related to surgical procedures. One of the challenges therefore is how this complex knowledge pattern can be unraveled and represented in a formalized way. We propose an approach to rubrics starting from Fillmore's frame theory (1976, 1982, 1994) which results in a frame-based representation format with slots and fillers to be described in the following section in greater detail.

3 Frames and rubrics

Frames are particularly applicable to the conceptual representation of specialized domains such as surgical procedures (cf. Faber et al. 2005, 2007). The frame model we developed for the conceptual representation of rubrics is in some ways similar to the FrameNet approach (2006). Such as in FrameNet we treat frames as script-like conceptual structures that describe a particular type of event evoked by lexical items. In the case of rubrics these frame-evoking lexical items are nouns derived from verbs denoting surgical procedures such as 'replacement', 'removal', 'insertion', etc. However, our approach

clearly differs from FrameNet. First of all, our aim is not to document the range of semantic and syntactic combinatory valences of surgical procedure terms in all possible senses in order to create a lexical database, which is the ultimate goal of FrameNet. Instead, starting from a closed set of rubrics we aim to formalize the conceptual meaning associated with surgical procedure terms as used in rubrics in order to support their correct interpretation. The proposed frame consists of a limited number of slots or features (comparable to the so-called frame elements in FrameNet) which specify the interrelations of the head procedure term and the concepts headed by this term. These interrelations are often implicit and not derivable from the surface structure of rubrics. As aforementioned, rubrics are designed for statistical purposes and hence represent a purpose-dependent selection of concepts, which are in some way or another associated with a surgical procedure, rather than a full description of a procedure. Therefore, the proposed frame format differs from what could be termed a 'canonical' procedure frame in order to be consistent with the claim that the frame-based representation of rubrics should reflect as clearly as possible the meaning of the described procedure. This implies also that certain slots which one anticipates for the representation of surgical procedures are lacking in our frame (for example, there is no actor slot in spite of the fact that each surgical procedure evokes an actor because this information is not relevant for the interpretation of rubrics).

Basically, the model starts from the assumption that frames are representing a coherent structure of related concepts which form a gestalt in the sense that they are interrelated such that without knowledge of all of them, one does not have complete knowledge of one of the either. Surgical procedures indeed can be conceptualized as a constellation of interrelated processes, actions and entities. The different concepts referred to by rubrics represent a number of macro-categories such as the main procedure and (sub)procedures, as well as other categories[2] such as body parts, instruments and pharmaceutical substances (cf. subsection 3.1). These interrelated conceptual categories act as templates for other concepts in the domain, as well as for the conceptual subframes. Moreover, these categories highlight the multidimensional nature of specialized concepts, which is a key aspect for their full description (Bowker 1993).

3.1 The design of the frame model

Our frame-based representation of rubrics starts from the presupposition that rubrics are particular constructs in which specific pieces of surgical procedures are linguistically realized by means of terms referring to a limited set of concepts such as surgical deeds, body parts, pathologies, etc. These terms evoke a frame of medical knowledge associated with the referent scene. The aim of the proposed frame representation of rubrics is to render by means of slots and fillers just those knowledge pieces which are necessary to correctly interpret the rubric's meaning[3]. Slots and fillers have to be situated on the conceptual level: slots are relevant aspects of concepts, and fillers are the specifications. This can be illustrated by means of the rubric 'Heart transplantation' (cf. Figure 2).

[2] In FrameNet terms these concepts represent 'core elements' (such as body part) as well as 'non-core' (such as purpose or diagnosis).

[3] So, for example, not included within this frame will be the day of the week on which an event occurred or the body mass index of the patient participating in the surgical event, even though in medical reality such factors are part of the event.

Medical procedure ***Heart transplantation*:**

SLOTS	**FILLERS**
consists_of	operative action(s)
has_object	heart *heart/cardiaque*
has_diagnosis	heart disease
has_purpose	to restore heart function
...	

Figure 2. Procedure Frame 'Heart transplantation'.

This medical procedure can – simplified – be characterized by the slots 'consists_of', 'has_object', 'has_diagnosis' and 'has_purpose' (cf. infra). In other words, the rubric 'Heart transplantation' implies the diagnosis of a specific heart disease and consists of a sequence of actions which are carried out on the heart. These slots are specified by means of fillers which are, in their turn, concepts of a given category. In the given example the filler 'heart' specifies the slot 'has_object'. Cross-linguistically the fillers may be realized in different ways. As the example shows, in English the noun 'heart' expresses the 'has_object' relation, whereas in French the same filler would be an adjective ('greffe cardiaque'). In other words, fillers may belong to different syntactic classes such as nouns, adjectives, bound morphemes, etc. However, it is possible to predetermine the conceptual category of fillers by means of restrictions. For example, if the procedure concept is an 'incision' the semantic relation 'has_object' is restricted to an anatomical entity such as a heart valve in order to make sure that the relation is meaningful in semantic terms. By contrast, a procedure such as 'injection' requires a filler of the category 'pharmaceutical substance' to express this relationship in a meaningful way. An additional important feature of the proposed frame is that fillers in their turn are concepts which can be specified by means of definitory subframes. Figure 3 shows the subframe associated with the concept 'mitral valve'.

Mitral Valve

isa	heart valve
is-part-of:	heart
has-location:	between left atrium and left ventricle
consists-of:	two triangular flaps, that open and close
has-function:	regulates blood flow between the left atrium and left ventricle

Figure 3. Definitory Subframe of the Concept 'Mitral Valve'.

In order to design such a generic language-independent format with slots and fillers for cardiovascular rubrics, first of all the conceptual structure of that type of rubrics has to be investigated in greater detail. For the purpose of this investigation, we analyzed a parallel corpus of rubrics (n=143) in English, Dutch, French and German all of which refer to procedures on the cardiovascular system. The corpus has been

drawn from the procedure volume of ICD-9-CM. Starting from the lexico-syntactic structures of these rubrics we identified in the four languages a restricted set of concepts belonging to different categories. For this categorization we adopted a definition of the notion ‘conceptual category’ inspired by prototype theory according to which concepts form a category on the basis of similarity. This means, for example, that also artificial body parts such as a mechanical heart valve or a prosthesis are considered to be members of the concept category ‘body part’ because they share not all, but specific features of the original body parts (e.g. the function) and have become an integral part of the human body. According to this definition the concepts in rubrics can be assigned to the following categories: (1) medical procedure, (2) main action, (3) body part, (4) medical instrument and (5) pharmaceutical substance.

An important note with respect to the frame design concerns the differentiation between the ‘medical procedure’ on the one hand and the ‘main action’ on the other. In the given context the category ‘medical procedure’ is the equivalent of a rubric, which is the description of a given procedure. By contrast, a ‘main action’ is not necessarily the equivalent of a ‘medical procedure’: a ‘main action’ invariably represents some concrete action, which is not the case for a ‘medical procedure’. For example, the rubric ‘Cardiotomy’ describes a not further specified medical procedure in which amongst other things an incision of the heart is made. Only this action is explicitly referred to by the term ‘-tomy’. But also the reverse is true. It is possible that a rubric refers to a procedure as a whole without explicitly referring to concrete actions. For example, the rubric ‘Heart transplantation’ refers to a procedure which consists of a series of actions such as ‘removal of the sick heart’ and ‘insertion of the donor hart’. The term ‘-transplantation’ implies all these different unspecified actions. In other words, rubrics presuppose pragmatic knowledge as they do not provide a full description of medical procedures. Instead, specific actions carried out during the procedure are selected or simply omitted.

The difference between both concept categories can also be described in syntactic terms. As already mentioned (cf. section 1.3) rubrics can be characterized as nominal constructions with an underlying predicate-argument-structure in which the predicate corresponds with a two-argument action verb. This verb may represent a single concrete action (such as ‘incision’, ‘removal’) or a number of not further specified actions (such as ‘correction’, ‘revision’). In the first case the predicate is equal to the medical procedure, in the second case not. Because of the fact that each medical procedure presupposes an object each procedure (or rubric) consists at least of two concepts: the predicate (which corresponds with the procedure or not) and a concept of the category ‘body part’ as the argument in object position. As the corpus analysis shows medical procedures may also have concepts of other categories in argument positions (e.g. ‘medical procedure’). These concepts are optional arguments and are comparable with the ‘non-core’ frame elements in FrameNet. Examples of medical procedures with the minimal set of concepts (i.e. the procedure with predicate function (underlined) and the body part with object function (in italics) are: *annul*oplasty, *heart* biopsy, *infundibul*ectomy, etc. An example of a medical procedure with different concept types in object position is ‘Infusion (medical action) of *thrombolytic substance* (pharmaceutical substance) in *coronary artery* (body part)’.

As already mentioned each procedure presupposes a number of concrete actions which, in turn, constitute a conceptual category. These actions can be subdivided into main actions (i.e. actions which are necessary to carry out the medical procedure described by the rubric) and subprocedures (i.e. actions carried out in order to support the main action). The rubric predicates may or may not correspond to the main action and/or subprocedure(s). For example, the rubric ‘Replacement of mitral valve with tissue graft’ refers to the operative repair of the mitral valve in which the insertion of the tissue graft is the main action, and the removal of the defect mitral valve is the preceding subprocedure. In this case, the predicate ‘replacement’ corresponds to neither of the two. Furthermore, it is possible that a rubric only refers to a subprocedure such as in ‘Cardiotomy’ (cf. supra). Finally, each medical procedure represents a specific type of action. For example, ‘-ectomy’ is an action which consists of the removal of something (e.g. tissue), an ‘anastomosis’ is an action in the course of which a connection is created (e.g. a bypass), etc. For analyzing and inventorying the different action types associated with cardiovascular procedures we have enlisted the assistance of domain specialists. It shows that the following set of basic action types can be identified, namely (1) insert, (2) incise/open, (3) massage, (4) reposition, (5) close, (6) transform and

(7) remove. In fact, the main challenge in constructing a procedure frame is to handle the inherent recursiveness of actions associated with cardiovascular procedures. As our main objective is to represent the meaning of a rubric as economic as possible, the subprocedure is treated as being a subframe of the main action. Both concepts are related by means of the slot 'is_associated_with'. In this way the recursiveness of the actions can be reduced to the subprocedure which is the most relevant in the given context. Because of the fact that each subprocedure is some action too, it shares the same features as the main action which means that the same slots can be used for its representation.

Summarizing, the following concept types are represented in our frame model: (1) medical procedure, which is subdivided into (2) main action and one or more (3) subprocedure(s), (4) diagnosis, (5) purpose, (6) body part, (7) medical instrument and (8) pharmaceutical substance. In order to determine the kind of relations between these concepts we consulted again domain specialists. On the basis of their pragmatic input we identified the following conceptual relations. The (1) medical procedure is interrelated with the (2) main action by the conceptual relation 'consists_of'. The main action, in turn, is related to the (3) subprocedure by the relation 'is_associated_with' which is, in fact, a rather vague labeling. In the given context this relation means that each main action evokes several subprocedures which are for economic reasons treated as subframes (cf. supra). Both the main action and the associated subprocedure(s) are related to the concepts (6) body part, (7) medical instrument and (8) pharmaceutical substance by the relations 'has_object' (e.g. body part, pharmaceutical substance) and ' makes_use_of' (e.g. medical instrument). These slots then are specified by fillers such as 'heart' for body part, 'catheter' for medical instrument, and 'therapeutic fluid' for pharmaceutical substance. Finally, the medical procedure is related to the concepts (4) diagnosis and (5) purpose by the relations 'has_diagnosis' and 'has_purpose'. Figure 4 illustrates how the conceptual structure of the investigated rubrics can be represented by one and the same medical procedure frame.

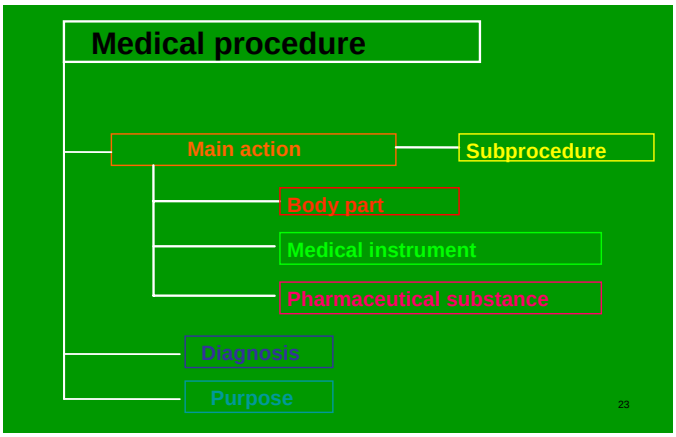


Figure 4. Example of the Procedure Frame.

3.2 The potentials of the frame model for rubrics

What are now the possibilities of this representation in practical terms? In our view, a frame-based format supports in a substantial way four things, namely the explicitation, the translation, the terminological standardization and the generation of rubrics. We also assume that the integration of the knowledge contained in the frame model into NLP software can be very useful. For example, a software program could, using the frame knowledge, extract a given rubric and its associated code from the operation file in a completely autonomous way. Potential users of this frame based software are in the first place the coders who have to interpret rubrics in order to code, but also the designers of rubrics (mostly medical

specialists), the translators of rubrics in the different target languages, and finally language processing machines. The following screen shots illustrate the design of our prototype editor for rubrics and the potentials of this software implementation by means of the rubric: ‘Closed heart valvotomy, aortic valve’ (cf. Figure 5). The main action in this rubric is ‘–tomy’ which belongs to the concept class ‘incision’/‘opening’. ‘Heart valve’ is the object of the action of ‘incision’.

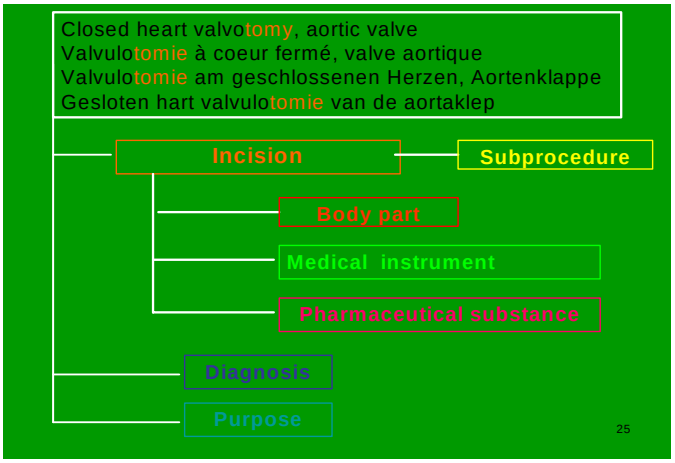


Figure 5. Frame Representation ‘Closed Heart Valvotomy, Aortic Valve’.

If the user clicks on ‘heart valve’ a pick list with all potential heart valves appears (cf. Figure 6).

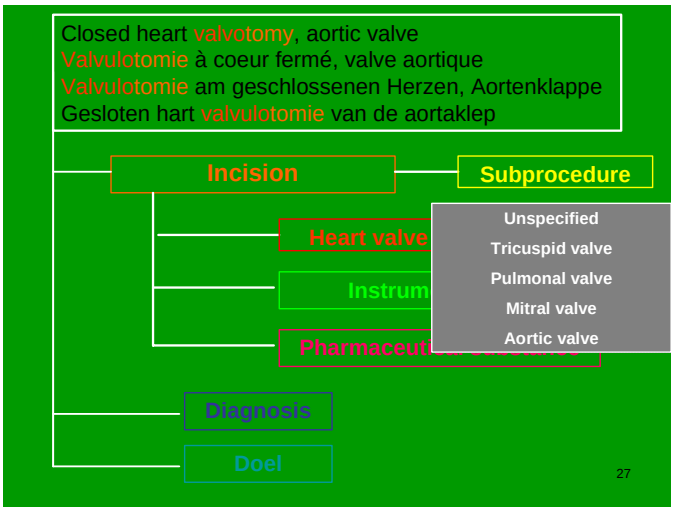


Figure 6. Pick List for ‘Heart Valve’.

As Figure 6 illustrates, the proposed frame representation supports the knowledge acquisition (that is the creation of a database), the definition of restrictions concerning the relations between the concepts and the restrictions concerning the specifications of the fillers. In the given case, ‘aortic valve’ is to be chosen from the pick list. An additional asset of this representation is that the different colors of the slot/filler-

combinations are also highlighted in the rubric itself. In this way it becomes immediately transparent how the conceptual structure of the rubric is syntactically and linguistically realized. Finally, it is possible (at least for the layman) to disambiguate the opaque phrase ‘Closed heart’ (cf. Figure 7). Obviously, the medical expert, who for example may use the given module for knowledge acquisition, knows that the term refers to a subprocedure in which the heart is left closed and a catheter is inserted through the skin (percutaneously) into a vein and pushed forward to the aortic valve in order to carry out the incision. Because the phrase ‘closed heart’ obviously refers to some procedure, it can be represented as ‘insertion of a catheter into the heart’ in the frame model. The conceptual structure of the term is explicitated by clicking on the string ‘Closed heart’. In addition, the expert can provide synonyms for the term ‘Closed heart’ in the database (like ‘Insertion of catheter into the heart’, ‘Percutanous technique’ or ‘Non-invasive technique’). Moreover, the editor shows preferred terms like ‘non-invasive technique’ in the given case. It goes without saying that the possibility to suggest preferred terms will generate an added value in terms of active terminological standardization of medical classification rubrics.

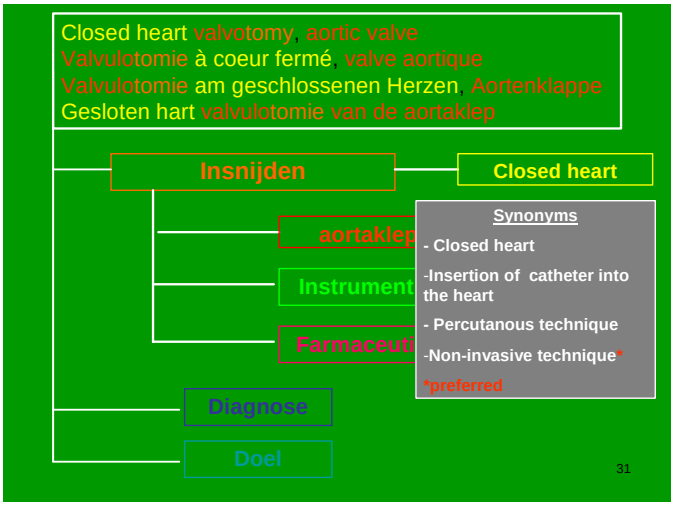


Figure 7. The Representation of Synonyms and Preferred Terms.

If we look at the final representation of the rubric ‘Closed heart valvotomy, aortic valve’ and its translations in Dutch, French and German (cf. Figure 8), the expressive advantages of this frame based software implementation in terms of disambiguation and terminological standardization are obvious.

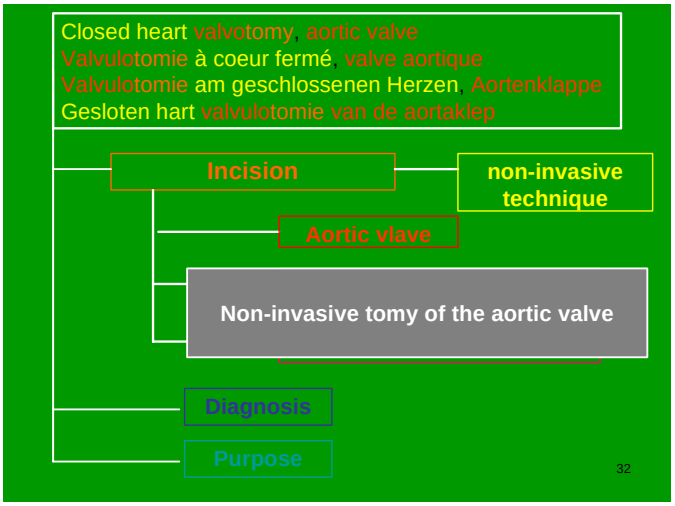


Figure 8. Representation ‘Closed Heart Valvotomy, Aortic Valve’.

First of all, the conceptual structure of this potentially ambiguous rubric is explicitated in a quite simple, but effective way both with respect to the medical expert and the layman. The difficult phrase ‘closed heart’ is paraphrased with the readily comprehensible phrase ‘Insertion of catheter into the heart’. Moreover, the relationships between the linguistic surface structures (i.e. the terms and phrases) and the corresponding concepts are visualized by means of different colors. In the same way one immediately can check whether or not the rubric is correctly translated, thus supporting the necessary control of translations. The editor also allows for the continual style improvement which keeps the legibility of rubrics in good stead. In the given example the phrase ‘heart valve’ is redundant as it occurs both in ‘valvotomy’ and in ‘aortic valve’. A final remark concerns the fillers of the slots ‘Diagnosis’ en ‘Purpose’. In the example at hand these slots are not specified, but this does not imply that this type of knowledge cannot be imported into the system. In order to make the system more powerful for potential end users the rubric designer easily can for example feed into the database all kinds of potential diagnoses causing the given procedure. In this case the database would display the filler terms ‘aorta valve sclerosis’ in combination with the associated ICD-9-code.

4 Conclusion

In this paper we take a closer look on medical classification rubrics which represent a highly specialized text type. Rubrics have a pragmatic function which is to convey information about codes in an efficient way. They are linguistically realized as more or less complex nominal strings consisting of medical terms which refer to a restricted set of concepts such as surgical actions, body parts, pathologies and medical substances. One of the main features of rubrics is their implicitness and potential ambiguity which may cause interpretation problems for both the professional coder and the translator of rubrics. As a matter of fact the structure of rubrics of different health domains already has been investigated in medical informatics which amongst others resulted in the CEN 1828 standard (1999). However, the approach adopted in this standard lacks a clear defined theoretical framework which results in inadequate descriptions of the conceptual categories and relations to be identified in rubrics. Therefore we decided to carry out a manual investigation of a restricted set of cardiovascular rubrics starting from a cognitive approach to meaning. This implies that the meaning of rubrics is not seen as equivalent to terminological meaning and the knowledge explicitly expressed in the surface structures. Rather, we proceed from the assumption that each single term in rubrics evokes a frame of conceptual knowledge relating to the specific scene referred to. In order to represent this knowledge in a formal way we developed with the support of domain specialists a frame-based representation of rubrics. This model is intended to make the conceptual structures of rubrics explicit and to support a number of applications such as the translation and generation of rubrics. Till now the usefulness of our approach for prospective users (i.e. professional coders) has not been tested. This is without any doubt an important task to do as only the effective use of the proposed frame model in daily practice will reveal its shortcomings. For the time being this case study is supposed to confirm the claim that specialized discourse as represented by rubrics may benefit from a frame-based modeling.

References

Bowker, Lynne and Ingrid Meyer 1993. Beyond ‘Textbook’ Concept Systems: Handling Multidimensionality in a New Generation of Term Banks. In K.D. Schmitz (ed.), TKE’93: Terminology and Knowledge Engineering. Frankfurt: Indeks Verlag, 123–137.

CEN EN 1828. 1999. Health Informatics. Categorial structure for classifications and coding systems of surgical procedures.

Faber, Pamela, Carlos Márquez, and Miguel Vega. 2005. Framing Terminology: A process-oriented approach. META 50 (4).

Faber, Pamela, Pilar León, Juan Antonio Prieto and Arianne Reimerink. 2007. “Linking images and words: The description of specialized concepts”. International Journal of Lexicography, 20 (1), 39–65.

Fauconnier, G. 1985. Mental Spaces: Aspects of meaning Construction in Natural Language. Cambridge, MA: MIT Press. Fillmore, Ch. (1985) “Frames and the semantics of understanding”. Quaderni di Semantica, 6(2): 222–254.

Fillmore, Ch. 1976. “Frame semantics and the nature of language”. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280: 20-32.

Fillmore, Ch. 1982. “Frame semantics”. In The Linguistic Society of Korea (ed.) Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co., 111-137.

Fillmore, Ch. And Atkins, B. 1994. “Starting where the dictionaries stop: The challenge for computational lexicography”. In B. Atkins and A. Zampolli (eds.) Computational Approaches to the Lexicon. Clarendon Press, Oxford: 349–393.

Gries, St. & Stefanowitsch, A. 2006. “Computational Approaches to the Lexicon”. In St. Gries, St. and A. Stefanowitsch (eds.) Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis. Berlin/New York: Mouton de Gruyter, 349-393.

ICD-9-CM. 2006. International Classification of Diseases, 9th revision, Clinical Modification, Hospital Edition. Los Angeles: PMIC.

Langacker, R. W. 1999. Grammar and Conceptualization. Berlin / New York: Mouton de Gruyter (Cognitive Linguistics Research, 14).

Langacker, R. W. 2001. “Discourse in Cognitive Grammar”. Cognitive Linguistics, 12, 2: 143-188.

L’Homme, Marie-Claude, Claudine Bodson and Renata Stela Valente. 1999. “Recherche terminographique semi-automatisée en veille terminologique: experimentation dans le domaine médical”, Terminologies nouvelles 20, 25–36.

Pearson, Jennifer. 1998. Terms in Context. Amsterdam/Philadelphia: John Benjamins.

Ruppenhofer, J. M. Ellsworth, M. Petruck, C. Johnson, J. Scheffczyk. 2006. Framenet II: Extended Theory and Practice. Berkeley: FrameNet Project.

Talmy, L. 2000. Toward a Cognitive Semantics. Cambridge, MA: MIT Press.

Wermuth, M.-C. 2005. Een frame-gebaseerde benadering van classificatierubrieken: Cardiovasculaire rubrieken als case study. Unpublished doctortal dissertation, Vrije Universiteit Amsterdam, Netherlands.

Index of authors