Workshop Programme

14:30	**Opening session**	
15:00	Mary Lourdes de Oliveira Angotti	*Terminological Variation and Conceptual Structure in Terms of Medicament Use Directions - Labels*
15:30	Sahara Iveth Carreño	*Characterizing Term Variation on an English-Spanish Parallel Corpus*
16:00	**Coffee break**	
16:30	Hee Sook Bae, Marie-Claude L'Homme & Guy Lapalme	*Semantic Roles in Multilingual Terminological Descriptions: Application to French and Korean Contexts*
17:00	Yukie Nakao	*Multilingual modalities for specialised languages*
17:30	Peter De Baer, Koen Kerremans & Rita Temmerman	*A Categorisation Framework Editor for Constructing Ontologically Underpinned Terminological Resources*
18:00	**Closing session**	

Organiser(s)

Béatrice Daille LINA-CNRS, Université de Nantes
Kyo Kageura, Library and Information Science, University of Tokyo
Marie-Claude L'Homme, Observatoire de linguistique Sens-Texte, Université de Montréal

Programme Committee

Table of Contents

Author Index

Terminological Variation and Conceptual Structure in Terms of Medicament Use Directions – Labels

Mary Lourdes de Oliveira Angotti
Universidade Federal do Triângulo Mineiro
Central de Idiomas Modernos – CIM
maryloa@terra.com.

Abstract

The aim of this work was to investigate the sociolinguistic aspects of the terms of medicament use directions. The *corpus* was comprised of 572 patient-addressed directions which have been made available on the Anvisa web site www.anvisa.gov.br. This publication is part of the project Anvisa *e-bulas* to simplify the language of medicine labels (instructions and information text on medicament use that goes inside the medicine box) in Brazil. In this sense, two texts were elaborated: one direct to the health specialist and another to the lay, medicine user. Regarding the terminological issues, three scientific levels were observed in the pacients' labels. First, the competitive variants Faulstchi (2001, 1999, 1998) are predominant in texts of the more specialized level. Second, in the more didactic level, the co-occurrent variants are predominant. Then, in the less specialized level, the concurrent variants are predominant. It was also found affixes and roots of classical origins (Latin and Greek) within the formation of the terminological units (UTs) predominant in more specialized terms.

1. Standardization of the medical terminology

In history, there have been controversies related to the standardization of the anatomical terms among health professionals. Every five years, there is a meeting with anatomists from many different countries in order to discuss and present suggestions to modify the used terms. As a result, it is elaborated the anatomical nomenclature *Nomina Anatomica.* So, there is a review in the old list by creating an up-to-date list every five years. According to Sager (1993:144), an anatomical classification is based on topographic principles as well as on twelve functional systems like musculoskeletal, respiratory, digestive. In this way, veins, arteries, muscles, nerves, among others, are identified themselves by their position or by their function. In the update chart of *Nomina Anatomica*, *Digestive system* and *Auditory Tube* are identified by their functions; however *Tonsila Palatina*, *Uterine Tube* and *Calcaneous Tendon* are identified by their position within the human body.

Diseases are classified according to their nature and origin: congenital, traumatic, infectious, neoplastic, metabolic, endocrine, allergic, psychiatric, iatrogenic and idiopathic. The causes of the diseases can be classified in etiological categories which refer to the organ or part of the body affected, Sager (1993:145).
Regarding to medicaments directions in Brazilian Portuguese (BP), the following suffixes were assigned by diseases and symptoms: **-ção**, **-éia** and **–mia**, for instance as in: infec**ção**, cefal**éia** and dislipide**mia** Sager (1993:145). For the exams and tests, the medicaments directions present the following examples of morphological derivation: ausculta**ção**, cateteriza**ção** and angio**grafia** . It states that the procedures present specific suffixes too, such as: laparo**tomia** and mastec**tomia.**
Because of the poorly-defined informational content, it is easy to understand the difficulty of simplifying these terms, as well as the occurrence of variations in the trans-codified equivalents. Still, the productivity of the eponyms is verified particularly because they form neologisms with derivative suffixes.
According to Carvalho *et al.* (2000:22), Brazil is placed fifth in the world in medicament consumption, with one drugstore for each 3.000 inhabitants, more than two times the number recommended by World Health Organization. Additionally, they state that this country is placed first in deaths by excessive use of medicine away from heath professionals' control. According to data from Fundação Osvaldo Cruz, 30% of 80.000 yearly deaths are caused the improper use of medications.

In the medication direction texts, there is recurrent use of terms that do not belong to the vocabulary of a lay person, in addition to the use of the passive voice, phrases without subject and long sentences; few objective sentences. All of these make inefficient the written communication of vital information to the users, especially regarding use restrictions, side effects and dosage.

For Cabré (1993) the specialized languages, when defined, should conciliate the theme criterion with the pragmatic conditions, such as the situation and the users. After analyzing fifty texts from ancient medicament direction versions, Angotti (2006 e 2007)

affirms that sixteen occurrences of synonym expressions were observed, which reveals a low number of clues available to the reader when inferring probable semantic contents belonging to his background knowledge.
RDC Resolution 140/2003 considers important, in addition to other factors, the size of the letter and the heterogeneity of information for the patient and for health professionals. Also, in order to make the language used in the medicament directions adequate to their standard users (non-specialized public) and health professionals (specialized public), this resolution established two types of medicament directions: one for the patient and another one for the health professional, according to what has been previously mentioned.

1.2 The process of linguistic alterations of medicaments directions

The simplification process presented in this work aimed at accommodating the language used in science to popular knowledge and language. Properties of the specialized knowledge texts were put together by many authors, among them, Sager (1993), Cabré (2001) and Lerat (1997). The linguistic aspects that the technicians responsible for writing the directions were orientated to alter were properties such as: the presence of complex morphology terms with formant of historical origin, the passive constructions and the nominalizations.

In 2003, during the process of adaptation to standards of medicament direction texts, and as a part of the Project "Medicament directions", elaborated by Anvisa, *two workshops were held with the representatives of the pharmaceutical companies*. The aim of these events was to present new linguistic characteristics to be adopted by the laboratories on direction texts for the medicament user. 135 medicament laboratory representatives, most of whom were pharmacists, attended these meetings and were responsible for writing the medicament directions.

With these workshops, it was intended to make the laboratory professionals aware of the importance of maintaining an efficient communication between the manufacturers and the users of the medicaments and motivate them to adopt the linguistic suggestions, which allowed simplification or vulgarization of the medicament directions language.

Pavel (2002:30) denominates this process as a terminological harmonization "which combines the desire for conceptual accuracy with linguistic correction". According to this author, the process of harmonization can be punctual or thematic, and it is conducted by a work-group or by a users' committee, which may or may not enlist specialists from the field in question.

Besides replacing the long sentences with shorter ones, the sentences in passive voice were rewritten, as well as nominalizations, so that the active voice and the sentence's subject became clear in the text. The SVC order was the one favored, and the use of the third person "the patient" was replaced by the second person "you". This way, a less formal and more direct communication with the medicament users was privileged, as shown below:

(1) In case of pregnancy, the patient should consult a doctor about the use of this medicament.

(1a) If you are pregnant, consult your doctor before using this medicament.

In the item labeled "indications", terms were found which allowed one to discern: (i) the action and indication of medicaments, (ii) the constitution of the medicament's composites and (iii) the way the patient should use the medicament. This part of the medicament directions text was used as a source of data for this study, that is, in order to compare the terminology used on the health professionals' medicament directions with the one used on the patients'.

Evaluation mechanisms are necessary with regard to the suppressions, adaptations and alterations of the patients' medicament directions' information in contrast with the professionals' directions, aiming at the maintenance of the information necessary to the lay reader.
It has been noted that the suffixes express the more general concept (conceptual extension) and have more generic meaning in the composite word, while the prefixes express more delimited notions (more specific) and relate to the intentional character of the concept.
In the item labeled "indications", terms were found which allowed one to discern: (i) the action and indication of medicaments, (ii) the constitution of the medicament's composites and (iii) the type of use the patient should make of this medicament. This part of the medicament directions text was used as a source of data for this study, that is, in order to compare the terminology used on the health professionals' medicament directions with the one used on the patients'.

In the patients' labels, three levels of language were used. These levels were classified according to the level of scientificity: more specialized texts, more didactic texts and less specialized texts.

The more specialized texts are the ones that show a predominance of scientific terms with classic origin (from medical terminology) and a predominance of passive voice (see examples 1, 1a and 2).

The second level, the more didactic texts, was characterized by the use of didactic language with

technical terms followed by its meaning, generally between blankets (example 3).

The third stage, in the place of scientific terms, popular terms were used to refer to diseases. Examples 2, 3 and 4 show respectively:

More specialized:

(2) Dermatoses Inflamatórias
Dermatosis Inflammatories

Didatic:

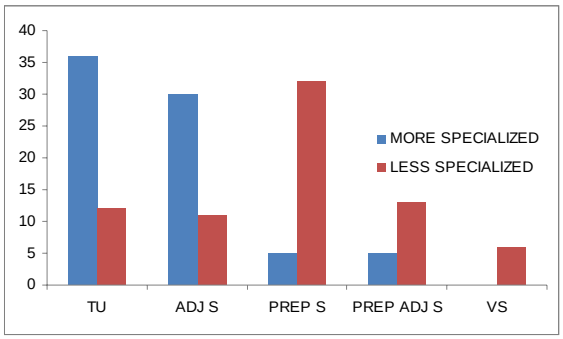
(3) Dermatoses Inflamatórias (doenças infecciosas da pele)
Dermatosis inflammatories (diseases infections of skin)

Less specialized:

(4) Micoses
Mycosis

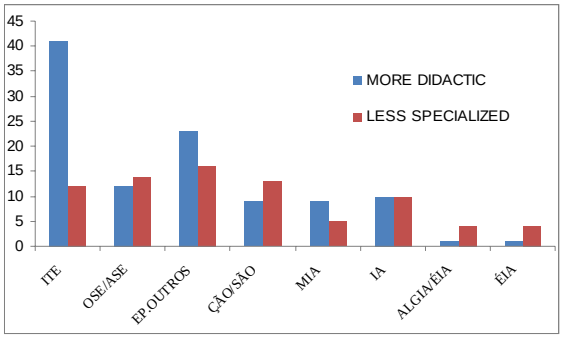
As for Galisson (1979:74-5), the lexical banalization is a socialized manifestation of the accomodation process that works over the bases of a large consensus. This manifestation has a stable format, used for begginers. In this sense, they are more didactic for presenting the scientific term and its concept between parentesis (exemple 2).
Loffer-Laurian (1963:10) sees the scientific vulgarization as a specialized and scientific knowledge spread for non-specialists. He excludes from this level of language the didactic activity for non-official instances. In this work, 'less specilized' are the popular expressions, found in the lay vocabulary, used to refer to diseases of symptoms. These expressions do not follow concept: they substitute scientific terms for popular ones
When denominating the diseases and symptoms in the labels analyzed, 86 didactic technical terms were used – repeated with additional information, mostly prepositional terminological syntagms (TS) between parentheses, while 46 terms were less specialized, that is, replaced by Prepositional Syntagms (PS), Adjectival Syntagms (AS), Prepositional Adjectival Syntagms (PAS) and Verbal Syntagms (VS), or by word belonging to the common lexis. The below examples illustrate the processes of didactic and less specialized.
More Didactic:
(5) Sinusitis (infection of the sinuses)
(6)Inflamatory dermatosis (skin disease) or (dermatological disease)
Less specilialized:
(7) Mycoses
The replacement of the Terminological Units (TUs) from health expert labels, more specialized texts, by TSs on the patients labels, less specialized texts, with a lower level of specialization, was characterized and quantified. The below graphic shows the most recurrent types of Syntagms, according to the level of specialization, that is, in the replacement of TUs by lexemes:

GRAPHIC 1 – Occurrence and typology of the TSs in more specialized and less specialized texts



Graphic 2 shows how the processes of more didactic and less specialized present themselves, according to the occurrence or absence of suffixes in the terms found in the patients' simplified medicament directions.

GRAPHIC 2- Less specialized and more didactic levels referring diseases and symptoms according suffix used.



The most recurrent suffix in the denomination of illnesses was **–ite** (fifty occurrences, forty more didactic and ten less specialized). The group of eponyms and other suffixes had a total of 39 terms, 23 of which were more didactic and 16 less specialized. Both in the group with eponyms and in the groups with the –ite and –mia suffixes (nine more didactic and five less specialized), the terms in the process of more didactic were more recurrent.
There is a similarity between the syntactic structures of levels more didactic and less specialized: in both levels it is possible to notice the use of prepositive syntagms, adjective syntagms and prepositive and adjective syntagms. These syntagms substitute more specialized terms formed, mainly, by classic origin suffixes.
What distinguishes these two great groups, besides the greater occurrence of lexical variants in less specialized directions, is the pragmatic result: the

more didactic directions contain both the technical term and its meaning (in parenthesis), which makes them more didactic, since the less specialized ones replace the technical terms with lexemes from the common language or with syntactic structures (prepositive and/or adjective TS).
These structures can refrain the specificity of the information contained in the more scientific term, which can also happen with the more didactic terms. For instance:

More didactic Term

(9) **Pyelo**nephritis: pyelonephritis (infecion of the kidneys)

The **pyel-** prefix means skin or membrane, which has not been taken into account in the explanation in parenthesis.

Less Specialized Term

(10) **Hyper**cholesterolemia type II a and II b = Changes in the cholesterol levels

In this case, the meaning of the **hyper-** prefix (excess) has not been taken into account in the less specialized version.
As to the terminological variants, in the more specialized level the terms present less variants than in stages of less scientificity. Also, only in this level do we witness the use of competitive variants.
The more didactic terms present a predominance of the use of prepositional TS in the additional informations and repeat both the TUs and the TS of the more specialized level. In some cases, there is loss if intentional information. The formal co-occurrent linguistic variant is predominant in the additional information.
The less specialized texts present lexemes that replace the more specialized terms, normally with the ellipsis of one of constitutive elements of these Terminologic units, which characterize the lexical linguistic variant[1] (cf. exemple 6). The replacing of these terms with TSs (prepositional and adjective) can also be witnessed, as shown in examples (7) and (8).
(11) Urticaria idiopática crônica: urticária
(11) Urticaria Idiopathic Chronic: urticaria
(12)Dermatose: doença da pele; doença dermatológica
(12)Dermatosis: disease of skin; disease dermatological
(13) Osteoporose: perda de massa óssea
(13) Osteoporosis: loss of mass bone
Finally, it is observed the existence of lexemes that replace the scientific terms in the denomination of diseases, as shown in the following examples:
(14) Tinea versicolor: pano branco

[1] To know in details the variants' tipology – in a socioterminologic perspective – please see the The Linguistics Variants Model proposed by Faulstich (2001, 1999 and 1998).

(14) Tinea versicolor: cloth white
(15) Ascite: barriga d'água
(15) Ascites: belly water
(16) Trato gastrointestinal: estômago e intestino
(16)Treat gastrointestinal: stomach and intestine.
The terms in prefixation and suffixation of a base normally suffer loss of the conceptual referred be the prefix. There is loss of conceptual in the information specificity, both in additional information of the more didactic terms and in the replacement with less specialized equivalents.
It has been noted that the greatest obstacle to spread this specific knowledge is related to the medical terminology and, therefore, a glossary has been created for the lay reader. This glossary contains a list of abreviations, a list of pictures and 1070 entries, taken from the texts of patients' medicament directions. This glossary is available at www.saudetodavida.com.

2. Semantic Traits and Syntatic Choices

In the formation of Complex Terminological Units (CTUs), it is possible to observe the influence of the semantic-conceptual properties in the syntatic structures that present themselves as Adjective Syntagms, or Prepositive Syntagms, depending on the type of restrictor they are comprised of; Sager (1990,1993), Dik (1997), L'Homme (2003) e Dowty (1991). In the examples below, the Noun plus Adjective form is preferred, in other words, they are used more frequently than the Syntagms Prepositives by speakers of Brazilian Portuguese (BP):
(17) Dor muscular, em vez de dor nos músculos
(17) Pain muscular, instead of pain in the muscles.

(18) Dor menstrual, em vez de dor da menstruação
(18) Pain menstrual, instead of pain from the menstruation.
When the affected body part is less delimited [-outlined], the speakers use Syntagms Adjectives more frequently, for exemple, muscular pain. Moreover, the [More Delimited] [More Outlined] trait of the arguments which complement these structures determines the notion of event or result Dik, based on the classification by Rijkoff (2002) - Colective Names; Singular/plural Names, Masses Names and Generic Names. It is proposed that, in the formation of these structures, the semantic properties of the [more outlined] and [more homogeneity] traits be combined. This reading of semantic properties favours the understanding that the outline and homogeneity traits influence the choice of a given syntatic category (adjective or preposition) with regards to medicaments' terms. This choice depends on semantic properties such as outline and homogeneity present on the argumentative structure of the more determined trait, the restrictive element.
To illustrate the syntatic differences of these Noums, note the nominalizations form Adjective or

Prepositive Syntagms according to the type of restrictor that comprises it. Look at the examples repeated below:

(17) Dor muscular, em vez de dor nos músculos

(17) Pain muscular, instead of pain in the muscles.

(18) Dor menstrual, em vez de dor da menstruação

(18) Pain menstrual, instead of pain from the menstruation.

When the body part is less delimited [less outlined], the use of the adjective is preferred over the use of the preposition, even if the preposition does not make non-gramatical nomalization. Moreover, the notion of event or result is especified by the more Delimited [more outlined] trait of the arguments that complement the nominalizations.

According to Alves (2006:112), regarding the notion of countability (singular-plural) being a characteristic of flexion more than a semantic class, and facing the examples given, it is proposed that the Nouns change the semantic properties of the following traits: [more outlined] and [more_homogeneity], established by Rijkoff (2002), according to the table below:

	+ HOMO GENEITY	- HOMO GENEITY
+ OUTLINE	+ outlined, + homogeneous Singular, plural N: intestine(s)	+ outlined, - homogeneous Colective Noum: organs
- OUTLINE	- outlined, + homogeneous Noum of masses: blood, water,fluid, plasma	- outlined, - homogeneous Generic Noum: all systems, endocrine system.

The reading of the nominal semantic properties favours the understanding that the outlined and homogeneity traits hold influence on the choice of a given syntatic category (adjective/preposition) of nominalizations.

3. Bibliographical References

ANGOTTI, M.L.O. (2007) Equivalência Conceitual na Terminologia dos Textos das Bulas de Medicamentos. Tese apresentada na Universidade Federal de Brasília, Unb, 150 págs.

ANGOTTI, M.L.O. (2006) Coesão referencial nos textos de bulas de medicamentos ensaio teórico-discursivo. *Lingüística: caminhos e descaminhos em perspectiva*. Orgs. Luiz Carlos Travaglia et al. Universidade Federal de Uberlândia.

ANVISA, (2003) Resolução – RDC n. 140, de 29 de maio de 2003. Diário Oficial da União de 02 de junho in http://www.anvisa.gov.br/legis/resol/2003/rdc/140_03rdc.htm

CABRÉ, M. T. (1993) *La terminología – Teoria, metodología, aplicaciones*. Barcelona: Editorial Antártida / Empúries.

CABRÉ, M. T. *et all*.(2001) Las características del conocimiento especializado y la relación com el conocimiento general. *La terminología Científico-técnica: reconocimiento, análisis y extracción de información formal e semántica.* IULATERM – Institut Universitari de lingüística Aplicada ed. Maria Teresa Cabré y Judith Felin, Barcelona.

CARVALHO, M.B., FERREIRA, L.M.A. & ONRICO, E.G.D. (2000). A relação entre as marcas textuais das bulas de medicamentos e a identidade do leitor. *XLVIII GEL*, UNESP, Assis.

FAULSTICH, E. Variantes Terminológicas: princípios de análise e método de recolha. *Actes Teflexions méthodologiques sur lê travail em terminologie et em trminotique dans lês langues latines*. Nice, Realiter / Université de Nice p. 15-20, 1996.

FAULSTICH, E. (2001) Aspectos de terminologia geral e terminologia variacionista. *Tradterm – Revista do Centro Interdepartamental de Tradução e Terminologia***.** FFLCG/USP vol. 7 p. 11-40, São Paulo.

FAULSTICH, E. (1999) Princípios formais e funcionais de variação em terminologia. *Seminário de Terminologia Teórica*, Ponencia presentada em el Seminário de Terminologia Teórica en Barcelona, 28-29 de janeiro de 1999.

FAULSTICH, E. (1998) Príncipes formels et fonctionnels de la variation em terminologie. *Terminology* , v. 5(1). Amsterdam. Philadelphia, John Benjamins, p. 93-106.

FEDERATIVE COMMITTEE ON ANATOMICAL TERMINOLOGY. (1998) - *Terminologia anatomica*. Stuttgart, Georg Thieme Verlag.

GALISSON, R. (1979) Le phenomene de banalisacion lexicale. Contribution methodologique à laproche des langues despecialité. Paris, Hachette, p. 75-9.

HENGEVELD, K., RIJKSHOFF, J. & SIEWIERSKA, A. (2005). Parts of speech and word order. *Journal of Linguistics*, 40.3.

INTERNATIONAL ANATOMICAL NOMENCLATURE COMMITTEE. - *Nomina Anatomica*. Edinburgh, Churchill Livingstone, 1989.

LERAT, Pierre. (1997) *Las Lenguas Especializadas* Trad. Albert Ribas. Editorial Ariel, S.A. 1ª. Ed., Barcelona.

MANUILA *et alli*,(2003) *Manuila Dicionário Médico*. Ed. Medsi, 9ª. Ed. Tradução e adaptação para a língua portuguesa Prof. Dr. Geraldo José Medeiros Fernandes.

PAVEL, S. e NOLET, D. (2000) *Manual de Terminologia*. Traduzido por Enilde Faulstich, Bureau de la Traduction, Translation Bureau, Public Works and Government Services, Canada.

SAGER, J. C.(1993) *Curso práctico sobre el procesamiento de la terminología*. Trad. Laura Chumillas Moya, Fundación Germán sánchez Ruipérz, Madrid, Pirámide.

RIJKHOFF, J. 2002. *The noun phrase*. New York : Oxford University Press.

.

Characterizing Term Variation on an English-Spanish Parallel Corpus

Sahara Iveth Carreño
Observatoire de linguistique Sens-Texte, Université de Montréal
C.P. 6128, succ. Centre-ville, Montréal (Québec), Canada, H3C 3J7
E-mail: si.carreno.cruz@umontreal.ca

Abstract

Recent studies have demonstrated that term variation is a frequent phenomenon in specialized texts. Most of these studies are descriptive; they base themselves upon the analysis of corpora in order to propose a characterization of term variation in a specialized domain. There are also works which exploit specific types of term variants with the aim of improving the results yielded by terminology-related applications and tools. These studies are usually carried out from a monolingual perspective. We consider however, that more research from a bilingual and multilingual perspective needs to be done. This is why we carried out an analysis of 90 terms in an English-Spanish aligned corpus made of original and translated documents (470 000 words per language). The main goal was to describe the different ways in which each English term was translated or represented in the Spanish texts, and then to present a typology of all these term representations, treating them as variants. Besides various types of denominative variants, we observed some other frequent phenomena that were grouped into two additional categories; i.e, syntactic variants involving semantic changes and diverse term representations that go beyond the morphological, lexical and syntactic transformations.

1. Introduction

During the last two decades, term variation has been the subject of several corpus-based studies, which have shown that this is quite a frequent phenomenon and plays a key role in various terminology-related applications.

Most of these researches have been carried out from a monolingual perspective, with the aim of describing the types of variants occurring in diverse specialized corpora. English and French (Daille et al. 1996; Jacquemin 1999) are the two languages which have been widely analyzed, although there are also some works for Catalan (Freixa 2002), Japanese (Yoshikane et al. 1999), and Spanish (Freixa 2001).

There are as well some application-oriented works which take into consideration a number of term variants for the development or improvement of NLP resources, *inter alia*, automatic term structuring (Bourigault and Jacquemin 1999; Daille 2003; SanJuan et al. 2005), machine translation (Carl et al. 2004), automatic term extraction (Daille 1994), text mining (Ibekwe-Sanjuan 1998), information retrieval (Ville-Ometz et al. 2007) and even for the study of concept evolution (Ibekwe-Sanjuan and SanJuan 2005; Picton 2007).

However, studies in a multilingual – or at least, bilingual – context are scarce. Carl et al. (2004), Daille et al. (1994) and Suárez de la Torre (2005) are three examples of contrastive studies.

Considering the frequency and the potential impact of term variation, we believe that more research in this context is needed.

This is why we decided to carry out an analysis of term variants on a bilingual corpus made of specialized texts. The main goal of the research was to characterize the variation phenomena that tend to occur in a context of translation. The idea was to observe the different ways in which a given set of English terms were represented in the Spanish texts, and then to present a typology of all these term representations, which would be considered as variants.

For the analysis, we based ourselves upon two different approaches of term variation that turned out to be complementary in our research.

At first, only the denominative variants were targeted. We greatly inspired by the definition and typology proposed by Freixa (2006). According to the author, denominative variation "*can be defined as the phenomenon in which one and the same concept has different denominations* (Freixa 2006:51)". This definition implies the semantic equivalence between a term and its variants; and it is basically restricted to lexicalized forms. Besides, Freixa's approach does not base upon a particular or preferred lexical form of a term in order to analyze variants.

Considering the fact that our analysis relies on the observation of terms and their different translations, and that not all the observed phenomena would fall into the denominative category, a second approach was included, the one introduced by Daille et al. (1996). We aim at giving a more integral description of variation in a bilingual corpus. In Daille's view, "*a variant of a term is an utterance which is semantically and conceptually related to an original concept* (Daille 2005:182)". Three salient aspects of this definition make it relevant in our context. Firstly, it implies not only the lexicalized denominations of terms, but also some other formal representations that are commonly found in specialized texts. Secondly, a variant is derived from a "base term" or "original term", usually the authorized or preferred form of a term appearing in terminological resources. Thirdly, the notion of variant is not restricted to forms holding a conceptual equivalence; in other words, there could be relatively short semantic distance between term and variants. In the following sections, we will make

reference to the key concepts of both approaches to describe our findings. It is nonetheless worth noting that, for practical reasons, in our context, we use *original term* when referring to the English term and *base term* to refer to the corresponding Spanish preferred equivalent.

The remainder of this paper is organized as follows. Section 2 describes the methodology adopted for the analysis of terms and their variants in our bilingual corpus. Section 3 presents the results of the analysis as well as the categorization of term variants deriving from the study. In section 4, we discuss the potential impact that these particular variants may have on some terminology-related applications. We then present our conclusion and the future work (Section 5).

2. Methodology

For the term variation analysis, an English-Spanish corpus was built. It is made of specialized texts taken from the environment domain, and they deal with industrial chemicals that threaten human health and the environment. The original texts are those written in English and the Spanish texts constitute their translations. All the documents were gathered from regional and international environmental organizations' websites. They are official treaties, conventions, scientific reports and regional case studies.

The corpus, which contains 37 pairs of texts – approximately 470 000 words per language –, was aligned at the sentence level using an automatic aligner, Logiterm[1].

The selection of the English terms to be analyzed was done in a two-stage process. First, an automatic term extractor called *TermoStat* (Drouin 2002) was applied to the English part of the corpus in order to obtain a list of candidate-terms. Then, from the list produced by *Termostat*, a total of 90 terms were selected. The selection was based on three main criteria: a) terms to be retained should belong to the targeted specialized domain; b) terms had to show a high frequency in the corpus; 5 occurrences was established as the baseline; and c) terms should also have a relevant distribution through the corpus, i.e., they should appear in at least 4 of the 37 documents.

Moreover, unlike many of the existing monolingual studies on term variation, we decided not to restrict our list to multi-word noun units, but to also include terms belonging to other parts of speech, namely verbs and adjectives, as well as one-word units and certain acronyms that are particularly frequent. The objective was to find out if simple terms and terms other than nouns could also vary. A list of 90 simple and complex terms was constituted.

Since the analysis focused on the discovery of variants as alternative means to translate terms, we deemed appropriate to define as point of departure a list of the corresponding equivalent terms in Spanish that would constitute the base terms from which variants would derive. We decided to take as Spanish base terms the preferred equivalents given in multilingual terminology resources (such as Termium®[2], IATE[3] and EEA[4] terminology banks). Table 1 presents examples of the selected terms, their frequency and distribution in the corpus, as well as their respective Spanish equivalent established as the base term.

Once the terms were selected, we proceeded to manually analyze them in the aligned corpus. To facilitate the reading of every context where terms appeared, we used a bilingual concordancer.

English term	POS	Freq	Distr	Spanish base term
abatement	n.	15	5	reducción
anthropogenic	adj.	149	17	antropógeno
bioconcentrate	v. intr.	29	6	bioconcentrarse
deposition	n.	166	15	deposición
environmental fate	n.	16	9	destino ambiental
grasshopper effect	n.	5	4	efecto saltamontes
greenhouse gas	n.	74	5	gas de efecto invernadero
landfill	n.	98	17	relleno sanitario
persistent organic pollutant	n.	113	17	contaminante orgánico persistente
release	n.	889	27	liberación
sewage sludge	n.	58	10	lodos de depuración
volatile organic compound	n.	6	4	compuesto orgánico volátil

Table 1: Examples of simple and complex terms selected for the analysis of variation

In a database, we noted all the different ways and forms in which each term was translated, the occurrences of each translation case, the number of documents in which these cases occurred, as well as other relevant linguistic information. Some examples of variants found in our bilingual corpus are given in Table 2

3. Results

All the Spanish equivalents found for each English term and previously stored in the database were divided into different categories. The classification was greatly based upon the typology used by Freixa (2002), focusing on the description of denominative variants. Thus, in the cases where the Spanish base term was not used as the equivalent of the corresponding English term, we observed the following denominative variants.

A. **Use of a synonym** ("lexical substitution").
Regarding simple terms, this implies that the base term was substituted by another different one (ex.1)[5].

[1] Terminotix Inc. http://www.terminotix.com/

[2] http://www.termiumplus.gc.ca/
[3] http://iate.europa.eu/iatediff/
[4] http://glossary.eea.europa.eu/
[5] The examples are presented in the following order: first the English term, then Spanish base term in parenthesis, and finally the variant, marked with an arrow.

(1) *emission* (*emisión*) → *liberación*

Regarding complex terms, the substitution could also be complete (ex. 2), or only partial, i.e., only the head (ex. 3) or the modifier (ex. 4) were the objet of the substitution.

(2) *hazardous waste* (*desechos peligrosos*) → *residuos peligrosos*

(3) *body burden (carga corporal)* → *acumulación corporal*

(4) *flue gas (gas de combustión)* → *gas de escape*

English term	Spanish base term	Variants	# different variants
Landfill	relleno sanitario	- vertedero de basuras ___ - terraplén - ellos - depósito de basura - vertedero - dichos rellenos - confinamiento - relleno - éste	10
monitoring	monitoreo	- supervisión - vigilancia - seguimiento - rastrear - monitorear - monitoring - control ___ - vigilar	9
adverse effect	efecto adverso	- transtorno - afectar negativamente - efecto negativo - efecto perjudicial - consecuencia perversa - repercusión nociva - efecto nocivo - efecto	8
industrial chemical	producto químico industrial	- químico industrial - sustancia industrial - de ellas - producto químico (o producto secundario) industrial - sustancia química artificial - sustancia química industrial - producto (y subproducto) químico industrial	7

Table 2: Examples of variants observed in the bilingual corpus for some of the selected terms.

B. **Abbreviation**.
One single case of abbreviation was observed for several terms: when the term was part of an acronym (ex.5).

(5) *persistent organic pollutant (contaminante orgánico persistente)* → *COP*

C. **Morphosyntactic variation – derivational morphology**.
Although this type of variant does not appear in Freixa's typology, we deemed appropriate to add to the denominative variants all the cases where there was a transformation of the base term's part of speech (POS). Several POS tranformations were involved: from an adjective into a noun (ex.6), from a noun into an adjective (ex.7), from a verb into a noun (ex.8), from a noun into a verb (ex.9), and even from an acronym into a noun (ex.10), and a noun into a prefix (ex.11).

(6) *coal-fired (carboeléctrico)* → *carboeléctrica*

(7) *vector control (control de vectores)* → *antivectorial*

(8) *biomagnify (biomagnificarse)* → *biomagnificación*

(9) *deposition (deposición)* → *depositarse*

(10) *PCB (BPC)* → *policlorobifenilos*

(11) *trace (traza)* → *micro-*

D. **Morphosyntactic variation without modification of the syntactic structure**.
In some cases, there was a modification of grammatical elements, namely the choice of prepositions and singular-plural change (ex.12); they did not involve any lexical or syntactic change.

(12) *long-range transport (transporte a gran distancia)* → *transporte de grandes distancias*

E. **Morphosyntactic variation implying a modification of the syntactic structure.**
All forms that differ from the base terms with respect to the number and order of words constituing the syntagmatic unit were grouped in this sub-category (ex. 13 and 14).

(13) *indoor (en interiores)* → *en espacios cerrados*

(14) *occupational exposure (exposición ocupacional)* → *exposición por actividad laboral*

F. **Reduction.**
In our corpus, complex terms are commonly the subject of reductions, namely ellipsis. The most frequent case observed was the reduction of the modifier (ex.15), but particular reductions of the head were also present (ex.16).

(15) *landfill (relleno sanitario)* → *relleno*

(16) *industrial chemical (producto químico industrial)* → *químico industrial*

G. **Transformation of simple – complex terms.**

In some cases, simple terms were substituted by an equivalent formed by more than one word; i.e, a complex term (ex.17). The opposite phenomenon was also observed, i.e, a one-word synonym replaced a multi-word base term (ex.18); although the latter case was rare.

(17) *congener (congénere) → compuesto análogo*

(18) *adverse effect (efecto adverso) → transtorno*

Reductions are not to be mistaken for transformations, since the earlier involve the omission of a lexical unit of the base term, while the latter constitute synonyms of the base terms; they only differ in terms of syntactic length.

When analyzing complex terms, we found out that some of the variations present in the translations did not fall into any of the categories proposed by Freixa. In other words, these cases could not be classified as denominative variants, because they involved significative semantic differences between original terms/base terms and variants. In fact, authors like Daille (1994), Daille et al. (1996), SanJuan et al. (2005), and Ville-Ometz et al. (2007) categorize them as syntactic variants and make a further analysis of sub-types. We decided to create a second category that would group the two basic morphosyntactic changes: insertions and coordinations.

H. **Insertions.**

While analyzing some contexts where our target terms appeared, we observed that their syntactic structure was interrupted by the insertion of two types of elements: a) adjectives or adverbs qualifying the term (ex.19); and b) the inclusion of adjectives, adverbs or other nouns pointing out to a more specific term (ex.20).

It is important to note that the insertion usually takes place only in the Spanish texts, due to the difference regarding the synactic construction between English and Spanish.

(19) *smart action plan (plan de acción) → plan inteligente de acción*

(20) *occupational exposure to mercury (exposición ocupacional) → exposición al mercurio en el lugar de trabajo*

I. **Coordinations.**
In some cases, our selected original term shared the modifier (ex.21) or the head (ex.22) with another term.

(21) *risk assessment and management [risk assessment + risk management](evaluación de riesgos) → evaluación y manejo de riesgos;*

(22) *municipal, medical and hazardous waste (municipal waste + medical waste + hazardous waste) → residuos médicos, municipales y peligrosos.*

From the semantic point of view, these syntactic phenomena are interesting since they often imply the co-occurrence of more specific terms in the corpus, sometimes even antonyms, which shows the semantic distance between them.

Other frequent phenomena were also observed in the corpus, but they did not fall into any of the denominative sub-categories, neither are they related to the syntactic category above presented. They were thus separated into a third category that we simply called "others".

J. **Anaphora.**
If the original term appeared more than twice in a sentence, it was usually translated for the first occurrence and anaphora were used for subsequent occurrences (ex. 25 and 26).

(23) *sinter plant (planta de sinterización) → estas instalaciones*

(24) *incinerator (incinerador) → ellos*

The use of anaphora seems to be a frequent stylistic resource in our specialised corpus.

K. **Paraphrases.**
In certain cases, simple and complex base terms were not present at all in the Spanish sentences; instead a paraphrase of the specialized concept was observed (ex.25 and 26).

(25) *human exposure (exposición humana) → [...] los seres humanos pueden estar expuestos...*

(26) *coal-fired (carbonífero) → [...] plantas que utilizan el carbón ...*

L. **Omissions**.
On a frequent basis, when an original term occurred more than twice in the same sentence or paragraph, from its second occurrence on, it was omitted in the Spanish translation (ex.27). We observed omissions both on simples and complex terms.

(27) *hazardous waste* (*desechos peligrosos*) → ___

[En] Several agreements address transboundary shipments of hazardous waste including PCBs. These agreements establish a framework for domestic regulation of transboundary shipments of hazardous waste and other waste, which recognizes the right of a country to ban the export or import hazardous waste and other waste and allows the transboundary movements...
[Sp] Diversos acuerdos tratan los embarques transfronterizos de desechos peligrosos, incluidos los BPC, y establecen un marco de trabajo para la regulación interna de desechos peligrosos o de otro tipo; estos acuerdos reconocen el derecho de un país para prohibir la exportación e importación y permitir los movimientos transfronterizos ....

M. **Terms appearing both in the English and Spanish.**
Since we are in a translation context, it was common to observe in the Spanish texts that the original English term (usually new, highly specialized terms, or terms that were part of a proper noun) was placed right after its Spanish equivalent (in most cases between parenthesis) (ex.28).

Category	Sub-category	# cases (/231 variants)	%	# terms (/45)	%
Denominative	**synonym** (lexical substitution)	93	40.2	38	84.4
	abbreviation - inclusion in an acronym	2	0.8	2	4.4
	morphosyntatic variant – **derivational morphology** - adjective – noun (1) - noun – adjective (9) - noun – verb (17) - verb – noun (7) - acronym – noun (1) - noun – prefix (2)	37	16	19	42.2
	morphosyntactic variant – **no modification** of the syntactic structure	0	0	0	0
	morphosyntactic variant – **modification** of the syntactic structure	11	4.8	6	13.3
	transformation of a simple term into a complex term	11	4.8	9	20
	reduction - Of the modifier (1)	1	0.4	1	2.2
Others	original term appearing in the translation	17	7.3	17	37.8
	omission	26	11.2	26	57.8
	anaphora	16	6.9	11	24.4
	verbal phrase	2	0.8	2	4.4
	paraphrase	5	2.1	3	6.7
	wrong equivalent (translation problem ?)	10	4.3	7	15.5

Table 3: Types of variants observed in simple terms.

Category	Sub-category	# cases (/230 variants)	%	# terms (/45)	%
Denominative	**synonym** - total lexical substitution (14) - substitution of the head (24) - substitution of the modifier (24)	62	26.9	32	71.1
	abbreviation - inclusion in an acronym	4	1.7	3	6.6
	morphosyntatic variant – **derivational morphology** - noun – adjective	1	0.4	1	2.2
	morphosyntactic variant – **no modification** of the syntactic structure	3	1.3	3	6.6
	morphosyntactic variant – **modification** of the syntactic structure	31	13.5	18	40
	transformation of a complex term into a simple term	2	0.9	2	4.4
	reduction - Of the head (4) - Of the modifier (2)	24	10.4	20	44.4
Morphosyntactic (implying semantic changes)	**insertion** - of qualifying units (17) - generating an antonym (1) - yielding a more specific term (22)	40	17.3	14	31.1
	coordination - of the head (9) - of the modifier (10)	19	8.3	9	20
Others	original term appearing in the translation	10	4.3	10	22.2
	omission	5	2.2	5	11.1
	anaphora	12	5.2	8	17.8
	verbal phrase	10	4.3	7	15.5
	paraphrase	1	0.4	1	2.2
	wrong equivalent (translation problem ?)	6	2.6	4	8.9

Table 4: Types of variants observed in complex terms

(28) *municipal solid waste (desechos sólidos municipales) :*
[En] ...Australian Municipal Solid Waste Program...
[Sp] ...Programa de Desechos Sólidos Urbanos de Australia (Australian Municipal Solid Waste Program).

N. **Wrong equivalent.**
There were some particular cases where the concept represented by the original term did not correspond to the concept represented by the Spanish variant (ex.29); in some other situations, the selected equivalent introduced a sort of ambiguity (ex.30). We presumed that they were errors introduced in the translations. Without further analyzing them, we decided to group them under a last sub-category.

(29) *deposition* (*deposición*) *eliminación* → ['disposal']

(30) *risk assessment* (*evaluación de riesgos*) *evaluación de riesgo* ['risky or dangerous evaluation']

Finally, it is important to mention that types of the variants that we have described sometimes occurred together; i.e., the same occurrence of a term was the subject of a combination of two or three variation cases. For the purposes of the present description, we deemed more appropriate to keep them separated.

Overall, the analysis of simple terms yielded a total of **321** different types of variants; while the complex terms resulted in a total of **230** different variants.

Tables 3 and 4 summarize the results of the term variation analysis carried out in our bilingual corpus. Table 3 corresponds to the simple terms and Table 4 to the complex terms. In both tables, the first and second columns list, respectively, the category and sub-categories of variants that we have introduced above. The third column indicates the number of different cases for each sub-category of variant found in our corpus[6]. The fourth column indicates the percentage that each sub-category represents over the total number of variants observed. Finally, the fifth and six columns show, respectively, the number of terms that were the subject of each particular type of variation, and their proportion over the 45 analyzed terms.

4. Discussion of the results

As Tables 3 and 4 show, both simple and complex terms were represented in the Spanish texts by a wide variety of variants. Regarding the total number of variants, we confirm that both types of terms yielded practically the same figures. What is surprising is the fact that for all of the single terms we observed at least one variant; which did not happen with the complex terms, since three of them were always translated by the same equivalent (e.g. climate change → cambio climático).

[6] Usually, the very same case of variant occurred more than once in the corpus.

The simple terms having the highest number of different variants are *disposal* (18), *trace* (13), *coal-fire*d (10) and *landfill* (10). On the other hand, terms like *methylmercury*, *capacitor* and *bioconcentrate* present only one variant. It should also be noted that among the most variable terms, we find some adjectives and verbs, which confirms that not only nominal units vary.

Regarding the types of variants, the use of a synonym (40, representing 40% over the total number of variants) and the POS transformations (representing 16%) were by far the most frequent denominative variants. From a total of 45 analysed terms, 38% present at least one case of substitution by a synonym. Of course, no syntactic variants (second category) implying semantic changes were observed in single terms. Omissions and term had also a significant frequency in these groups of terms.

With respect to complex terms, a total of 230 different variants were found in our bilingual corpus. The terms that presented the highest number of variants were *sound management* (16), *hazardous waste* (14), *risk assessment* (13) and *action plan* (11); contrary to our presumptions, most of the three-word terms and some two-word terms presented a low variation rate (e.g., *volatile organic compound* (1) *persistent organic pollutant* (1), *hazardous air pollutant* (1), *fatty tissue* (1)).

Concerning types of variants, synonyms, morphosyntactic variants implying a modification of the syntactic structure as well as reductions were the most frequent ones; between 18 % and 32 % of the terms (over a total of 45) were the subject of these variants. As it was expected, insertions were the most frequent type of syntactic variants. From the "others" category, again, anaphora and the original term also appearing in the translation are two relatively frequent variants.

Taking into consideration the total number of variants (231) over the total number of analysed terms (45), we found out that the variation rate for both simple and complex terms is of **5,1**.

As we already mentioned, our initial goal was the characterization of denominative variation. Denominative variants are actually the most frequent types according to or analysis. Overall, they represent approximately 40% of the variants found in simple and complex terms.

However, other variation phenomena have proven to be significantly frequent, particularly syntactic variants, anaphora and omissions. In the following section we will explain some reasons why we emphasize these particular variants.

5. Potential impact of term variation on terminology-related applications

While there is no consensus when considering the non-denominative included in our analysis as being real variants of terms, we deemed important to describe them in a contrastive analysis for three reasons.

Firstly, they had a significant frequency in the analysis corpus. Secondly, it seems that a bilingual analysis,

especially when it is based on aligned translation corpora, makes these types of changes more easily observable. As a matter of fact, and to the best of our knowledge, monolingual descriptive studies do not take them into consideration. And thirdly, we think that some of these transformations still constitute, to a certain extent, non-conventional linguistic representations of specialized concepts.

Moreover, all these linguistic transformations could play an important role in terminology-oriented applications and even Natural Languages Processing (NLP) activities.

For example, in the context of computer-aided translation (CAT), the fact that the original terms rarely have a single equivalent could negatively affect the performance of translation memories and terminology translation programs (i.e., number of exact and fuzzy matches proposed by the tools). At the same time, some characterized variants might be modelled, so that they could be integrated by such tools and thus improve them.

As Carl et al. (2004) have shown, machine translation can also benefit from the modelling of specific denominative and syntactic variants in order to improve bilingual dictionaries. Some partial or total lexical substitutions, insertions, omissions and coordinations could be automatically adbduced from parallel corpora by establishing morphosyntactic transformation patterns.

Based on the principle that terms and their variants hold strong semantic links, automatic and semi-automatic terminology extractors have started taking into consideration term variants at different stages. For example, some systems conflate various morphosyntactic forms of a term appearing in texts into a unique canonical form at an early stage (Daille 1994; Nenadic et al. 2004). The goal is to count them as a single candidate-term rather than different independent ones. Other methods, on the contrary, rely first on the manual analysis of reccurrent paradigmatic and syntagmatic semantic relations that a given couple of term-variant holds (e.g., coordinations tend to involve co-hyponyms, certain insertions imply a hyperonym-hyponym relation) in order to structure candidate-terms (e.g., Bourigault and Jacquemin 1999). These strategies render automatic term extraction tools more attractive for human users, who prefer to see relevant candidates grouped over a simple alphabetically-ordered list.

The performance of automatic bilingual term extractors is often compromised due the various equivalents used for a single original term. It has been shown that anaphora, occurrence of the original term in the translation, insertion of qualifiers, POS changes and paraphrases negatively affect the term-equivalent alignment or matching. Take for example, the following aligned sentences from which a term and its equivalent would be proposed:

[En] Dietary **human exposure** through the food chain has been little explored.
[Sp] Se ha estudiado poco la **exposicion** relativa a la dieta de **seres humanos** a través de la cadena alimentaria.

While it also depends on the strategy applied, there are few probabilities that the system succeeds in proposing the right equivalent for the original term (in bold in the example).

In the fields of information retrieval, text mining, and ontology building, syntactic variants of terms have been exploited. Ibekwe-Sanjuan (1998), for example, developped an automatic method to extract a set of terms, recognize three types of their syntactic variants and then organize topics according to the semantic relations established between the terms and those variants.

One of the first possible limitations to implement such an strategy in a bilingual context is that, as our analysis has shown, most syntactic transformations only take place in the translation (smart *action plan* – regional *de acción*). Thus we think that it will be necessary to first establish common transformation models from one language into another one.

In order to further exploit term variants and, at the same time, overcome the limitations that term variation phenomenon imposes to different terminology and NLP fields of application, more linguistic analysis on semantic relations between variants in different languages is deemed appropriate.

6. Conclusion

The great variety of denominative, syntactic and other textual variants found in our parallel corpus shows, once again, that terminology variation is an important and frequent linguistic phenomenon in a translation context.

Contrary to what we initially supposed, simple terms and complex terms show a very similar tendency to vary, regardless of their parts of speech. Moreover, they both are subject to almost the same types of variations.

From a more general perspective, denominative variants were the most frequent in our English-Spanish corpus. Among them, the use of synonyms, abbreviated forms and morphosyntactic variants implying a modification of the syntactic structure were the most recurrent cases.

Encouraged by the results of our analysis, and recognizing that it was carried out only in one direction (from English into Spanish), we plan to expand our study by analyzing three different languages and bigger corpora. The objective is to compare the results in all the selected languages, and find out if the variation tendency (and rate) changes. We deemed appropriate to base this future analysis on a comparable corpora, since the frequency and types of variants may also change.

7. References

Bourigault, D. & Jacquemin, C. (1999). "Term extraction and term clustering: An integrated platform for computer-aided terminology", in *Proceedings of Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, pp. 15-22.

Carl, M., Rascu, E., Haller, J. & Langlais, P. (2004). Abducing term variant translations in aligned texts. Terminology, 10(1), pp. 103-133.

Daille, B. (1994). Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistiques. Thèse de Doctorat, Paris: Université Paris-7.

Daille, B. (2003). Conceptual Structuring through Term Variations. In F. Bond, A. Korhonen, D. MacCarthy & A. Villacicencio (Eds.), Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pp. 9-16.

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

Daille, B., Gaussier, E. & Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), Kyoto, pp. 515-521.

Daille, B., Habert, B., Jacquemin, C. & Royauté, J. (1996). Empirical observation of term variations and principles for their description. Terminology, 3(2), pp. 197-257.

Drouin, P. (2002). Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés. Thèse de Doctorat, Montréal: Université de Montréal, 274 p.

Freixa, J. (2001). Reconocimiento de unidades denominativas: incidencia de la variación en el reconocimiento de las unidades terminológicas. In M.T. Cabré, and J. Feliu (Eds.), *La terminología, científico técnica: Reconocimiento, análisis y extracción de información formal y semántica*, Barcelona: IULATERM, pp. 57-65.

Freixa, J. (2002). *Anàlisi de la variació denomitiva en textos de different grau d'especialització de l'àrea de medi ambient.* Thèse de Doctorat, Barcelona: University of Barcelona, 397 p.

Freixa, J. (2006). Causes of denominative variation in terminology. *Terminology*, 12(1), pp. 51-77.

Ibekwe-Sanjuan, F. (1998). Terminological variation, a means of identifying research topics from texts. In *Proceedings of the Joint International Conference on Computational Linguistics* (ACL-COLING'98), Montréal, pp. 654-570.

Ibekwe-SanJuan, F. & SanJuan, E. (2004). Mining textual data through term variant clustering: the Termwatch system. In *Actes de Recherche d'Information assistée par ordinateur* (RIAO), Avignon, France, pp. 487-503.

Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th Meeting of the Association for Computational Linguistics*, Maryland, pp. 341-348.

Nenadic, G., Ananiadou, S. & McNaught, J. (2004). Enhancing Automatic Term Recognition through Term Variation. In *Proceedings of 20th International Conference on Computational Linguistics* (Coling 2004), Geneva.

Picton, A. (2007). Repérer l'évolution des connaissances dans des textes spécialisés : Quelle(s) méthode(s)? Quels indices linguistiques? Quelles applications? In *Séminaires du Centre de Recherche en Terminologie et Traduction*, Université Lyon 2, January 31 2007.

SanJuan, E., Dowdall, J., Ibekwe-SanJuan, F. & Rinaldi, F. (2005). A symbolic approach to automatic multiword term structuring. Computer Speech and Language (CSL), Special Issue on Multiword Expressions, Elsevier, London, pp. 20.

Suárez De La Torre, M. (2005). Análisis contrastivo de la variación denominativa en textos especializados: del texto original al texto meta. Thèse de Doctorat, Barcelona: Universidad de Barcelona.

Ville-Ometz, F., Royauté, J., & Zasadzinski, A. (2007). Enhancing in automatic recognition of term variants with linguistic features. Terminology 13(1) pp. 35-59.

Yoshikane, F., Tsuji, K., Kageura, K. & Jacquemin, C. (1999). Detecting Japanese Term Variation in Textual Corpus. In 4th International Workshop on Information Retrieval with Asian Languages, Taipei, pp. 97-108.

Semantic Roles in Multilingual Terminological Descriptions: Application to French and Korean Contexts

Hee Sook Bae*, Marie-Claude L'Homme* and Guy Lapalme**

*Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec)
hee.sook.bae@umontreal.ca; mc.lhomme@umontreal.ca

** Recherche appliquée en linguistique informatique (RALI)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec)
lapalme@iro.umontreal.ca

Abstract

This article presents a project that aims at enriching a domain specific lexical database with a module in which contexts are annotated. The lexical database which contains French terms that belong to the domains of computing and the Internet is currently being extended to other languages, namely English and Korean. The annotation model (based on that used in FrameNet, Ruppenhofer et al. 2002) first takes into account the semantic roles of participants (e.g., AGENT, PATIENT, INSTRUMENT) which are assumed to be language-independent. The second part of the annotation describes the syntactic behavior of participants and thus also takes into account linguistic specificities. In this part, phrases are annotated according to their function and their type. We also encode prepositions used in French and cases and particles used in Korean. This annotation allows us to inventory syntactic information along with semantic roles. As a result, we can verify that the annotation model in French also applies to other different languages, even very different ones, such as Korean.

1. Introduction

A specialized lexical database – which contains terms related to the fields of computing and the Internet – is currently being extended to other languages, namely English and Korean (a Spanish version is also planned in the near future). The original database, called *DiCoInfo. Dictionnaire fondamental de l'informatique et de l'internet* (the French version is available at http://olst.ling.umontreal.ca/dicoinfo/) provides detailed information on the lexico-semantic properties of terms and contains five main sections: headword (along with grammatical information), actantial (i.e., argument) structure with lists of linguistically realized terms, lexico-semantically related terms for each headword, contexts and definitions. The English and Korean counterparts also contain rich linguistic information, but definitions have not been included yet.

The extension of our descriptions to other languages raises the question of ensuring that similarities between senses are represented using the same apparatus while taking into account linguistic specificities. To this end, we developed a method for annotating contexts in which terms appear that describes the semantic roles of participants, but also their syntactic behavior. We also designed an XML structure that is flexible enough to accommodate annotations in different languages. We are currently testing this method and the XML structure on French and Korean, two typologically different languages. This paper presents our method and its relevance in multilingual terminology descriptions that aim at describing the linguistic properties of terms.

This paper is structured as follows. Section 2 describes the general hypotheses underlying the project. Section 3 explains the basic principles on which our annotation model relies and the XML structure used for carrying out the annotations. The methodology devised to annotate French and Korean contexts is detailed in Section 4. Finally, Section 5 makes observations on differences between French and Korean that were encountered during the annotation process and some adaptations that were made to the XML structure.

2. Why semantic roles in a specialized lexical database?

Semantic roles are designed to capture the semantic relationships between a predicate and its actants (i.e., arguments). The specification of actants in terms of roles is an efficient and elegant way to use the same labels to refer to similar relationships shared by an actant and a

predicate as shown in (1) and (2) (cf. Baker et al. 1998; Fillmore 1968; FrameNet 2008; VerbNet 2008).

(1a) PATIENT (e.g., *a computer, a server*) starts
(1b) AGENT (e.g., *a user*) starts PATIENT (e.g., *a computer, a server*)
(2a) INSTRUMENT (*a printer*) prints PATIENT (e.g., *a file, data*)
(2b) AGENT (e.g., *a user*) prints PATIENT (e.g., *a file, data*) with INSTRUMENT (*a printer*)
(2c) printing of PATIENT (e.g., *a file, data*) with INSTRUMENT (a printer) by AGENT (e.g., *a user*)
(2d) printer used by AGENT (e.g., *a user*) to act on PATIENT (e.g., *a file, data*)

Semantic roles also appear to be an efficient apparatus to capture semantic similarities between languages regardless of surface phenomena which can set them apart.[1] For instance, *print*, *인쇄하다* and *imprimer* (3) have the same actantial structures, even if syntactic phenomena can differ.

(3a) print 1b
AGENT prints PATIENT with INSTRUMENT
(3b) 인쇄하다 1b
AGENT-이 PATIENT-을 INSTRUMENT-로 인쇄하다
(3c) imprimer 1b
AGENT imprime PATIENT avec INSTRUMENT

As can be seen in example (3), terms in all three languages have the same number of actants and these all have the same semantic roles. Terms differ here according to the preposition used as far as English and French are concerned (*with* and *avec*). In Korean, word order changes drastically and a case particle is added to the linguistic units that express a given actant: -이 for the AGENT, -을 for the PATIENT, and finally –로 for the INSTRUMENT.

3. A model for describing semantic roles

We developed a method (based on that used in FrameNet, Ruppenhofer et al. 2006) which consists in annotating the linguistic realizations of terms together with their participants (these include, but are not limited to, actants).[2] This annotation is performed on a number of sentences extracted from corpora. In addition to giving information on linguistic realizations of terms and their environment in real contexts, these annotations provide empirical evidence that support the terminologists' intuitions while writing entries in the lexical database.

The annotation model first describes the semantic roles of participants which are assumed to be language-independent. The second part of the annotation takes into account linguistic specificities and describes the syntactic behavior of participants. In this part, phrases and clauses are annotated according to their function and their type. We also encode all relevant information related to syntactic realizations (i.e.., prepositions used in French and cases and particles used in Korean).

3.1 XML annotation structure

The annotation is performed in an XML structure in which language independent abstract categories have been defined in order to accommodate the specificities of different languages. The annotation model has been designed with a view to serve the specific purposes of this project. Thus, no effort has been devoted up to now to adhere to annotations standards used to encode terminological data. The aim in our project is to annotate a significant number of contexts that will then be used to train a machine learning algorithm.

XML tags allow us to integrate the semantic role annotations within the contexts themselves. Appropriate tools can then be used to display them in different forms such as HTML pages (an example of annotated French contexts and the HTML page generated from these is reproduced in Appendix A). But the most important advantage of using XML is the fact that the annotations of the contexts can be validated against a schema (a grammar-like formalism illustrated in Appendix B) that defines the allowable tags and their embedding at each step of annotations.

The most important components of our annotation for the purpose of this article is the "participant" with a "type" ("Act" or "Circ") and a "role" (e.g., "Agent", "Patient", Instrument", "Manner" ...).[3] A "participant" contains a "syntactic function" with its name ("Subject", "Object", "Complement", "Modifier" ...) and – for languages in which this attribute applies – case ("Agentive", "Accusative", "Genitive" ...). A "syntactic function" contains more specific linguistic information such as "syntactic type" that will be presented further in the paper.

With these definitions, a schema aware XML Editor such as oXygen or XMLSpy) can be used to drive the

[1] Interestingly, FrameNet is also being extended to other non-Indo-European languages; e.g., Japanese (Kyoko et al. 2004; 2006) and Chinese (You et al. 2007).

[2] In addition to the several projects using the FrameNet framework to annotate lexical units in different languages, some extensions have been or are being developed to apply FrameNet-like descriptions to specialized languages, namely the Kicktionary (http://www.kicktionary.de/; Schmidt, forthcoming) and BioFrameNet (Dolbey et al. 2006).

[3] For the purpose of this article, the names of tags and attributes have been translated. They are given in French in the original version of the schema.

annotations in real-time showing at each step a menu of allowable choices at a certain point.

4. Methodology

The annotation of contexts itself is performed by terminologists and consists in four steps: 1) selection of terms; 2) definition of relevant semantic roles corresponding to realizations in contexts; 3) description of syntactic information (function and type); 4) use of the annotation model for contexts in different languages (and potentially adaptation to the specificities of these languages). The first three steps are described in the following subsections. The linguistic specificities related to French and Korean are discussed in Section 5.

4.1 Selection of terms

The annotation is carried out for predicative terms, in other words, terms having at least one semantic actant and a corresponding surface realization. Considering that verbs are the main predicates and that participants for this part of speech are most likely to be linguistically realized, we started with verbal terms. For each term, we annotate up to 20 contexts extracted from the corpora used at all stages of the work.

4.2 Definition of semantic roles

In its current state, our annotations already use around 15 semantics roles for actants. Among these, there are typical roles, such as AGENT, PATIENT, INSTRUMENT, DESTINATION.[4]

Figure 1 shows the application of our grammar to a sentence containing the verb *print*, namely *You can print the HowStuffWorks Big List of Computer Memory Terms*. The portion of the schema concerned with this example appears between lines 21 and 30 in Appendix B.

```
<participant type="Act" role="Agent">
… You …
</participant>
can
<lexie-att>print</lexie-att>
<participant type="Act" role="Patient">
… the HowStuffWorks Big List of Computer Memory Terms
…
</participant>
```

Figure 1: XML annotation of arguments

We also consider circumstants[5] in our annotations, and need a much larger set of semantic roles. Also, new roles must be defined, such as MANNER, PURPOSE, TIME. In this case, the type attribute will contain the value "Circ". Again, the portion of the schema concerned with this example appears between lines 21 and 30 in Appendix B.

4.3 Syntactic function

All phrases and clauses linked to each predicative unit are first analyzed in terms of their syntactic function (i.e., subject, object, complement, modifier, etc.). Figure 2 shows how syntactic functions are encoded in our sample sentence containing the verb *print*. The portion of the schema that corresponds to information appears between lines 40 and 44 in Appendix B.

```
<participant …> …
   <syntactic-function name="Subject">
   … You … </syntactic-function>
</participant>
can
<lexie-att>print</lexie-att>
<participant …>
   <syntactic-function name="Object">
   … the HowStuffWorks Big List of Computer Memory
   Terms …
   </syntactic-function>
</participant>
```

Figure 2: Annotation of syntactic functions

4.3 Syntactic types

Then, phrases and clauses are encoded according to their type (noun phrase, prepositional phrase, clause, etc.).

Figure 3 shows the model used for annotating syntactic types and its application to our sample sentence. The portion of the schema concerned with this example appears between lines 45 and 50 in Appendix B.

```
<participant …> …
   <syntactic-type name="NP">
   … You …
   </syntactic-type>
</participant>
can
<lexie-att>print</lexie-att>
<participant …> …
   <syntactic-type name="NP">
   … the HowStuffWorks Big List of Computer Memory
   Terms
   … </syntactic-type >
</participant>
```

Figure 3: Annotation of syntactic types

5. Adaptation of the annotation model to different languages

The same annotation method and XML structure are applied to predicative units in different languages, French and Korean.

Lexical units that correspond to terms are chosen in both languages according to the same methodology. A list a specific units is produced using the TermoStat (Drouin 2003) term extractor and then validated by

[4] Note that some roles used in our database differ from frame elements in FrameNet (2008). We try to define very general roles that can apply to long lists of terms and not only to units that appear in a specific frame.

[5] In FrameNet, these are called *non-core frame elements*.

terminologists applying four different lexico-semantic criteria (Bae & L'Homme 2006). The term extractor produces valid terms that belong to the following parts of speech: noun, verb, adjective, and adverb. As was said above, in this project we first focus on verbal terms.

Since lists of terms are validated independently in each language, we do not necessarily annotate equivalent verbs. Examples of French verbs are: *configurer* (Engl. configure), *accéder* (Engl. access), *partager* (Engl. share), *partitionner* (Engl. partition), *télécharger* (Engl. download). Examples of Korean verbs are: *공격하다* (Engl. attack), *공유하다* (Engl. share), *검색하다* (Engl. search).

However, French and Korean belong to completely different language families: the first is an Indo-European language and the second is an Altaic language. As an agglutinative language, Korean nouns accompany particles and verbs endings. In Korean, particles express cases and this allows words to be ordered freely in sentences, as compared with French or English. This section discusses some cases where differences were observed between the two languages and how our annotations were carried out in accordance with these.

5.1 Syntactic realizations of predicative terms and their participants

Although predicative terms have the same actantial structure in French and Korean, these can behave very differently when syntax is considered.

We will illustrate this with an example. The French verb *partager* and its Korean equivalent *공유하다* (Eng. share) have the same actantial structure (their participants are labelled as AGENT and PATIENT). However, in French, *partager* is transitive; the PATIENT role is expressed in the form of object noun phrases; AGENTS (A and B) can be expressed in different syntactic structures, as shown in (4a). In Korean, *공유하다* is also transitive; the PATIENT is expressed in the form of object noun phrases; AGENTS can be expressed in three different syntactic structures as shown in (4b).

(4a) … *partager les ressources* (PATIENT) *entre plusieurs utilisateurs* (AGENTS)
Un ordinateur (AGENT-A) relié *à une imprimante* (PATIENT) pourra *donc éventuellement la* (PATIENT) *partager*.
De nombreux professeurs (AGENTS) *s'en servent pour partager entre eux* (AGENTS) *des informations* (PATIENT), *des documents* (PATIENT), *les sujets donnés à leurs élèves respectifs* (PATIENT), *proposer des séquences de travaux pratiques, etc.*

(4b) *수십 개의 서비스가* (AGENTS) *하나의 인증 시스템을* (PATIENT) *공유하다.*
메인 사이트와 (AGENT-A) *서브 사이트가* (AGENT-B) *인터페이스를* (PATIENT) *공유한다.*
인터넷 연결 공유 서비스는 (AGENT) *2000 서버를 라우터로 사용하여 인터넷을* (PATIENT) *공유하고*..

5.2 Word order

In French, syntactic roles such as subject and object are decided by the order of phrases. In contrast, in Korean the subject can be positioned anywhere in the sentence. It is sufficient to find the nominal phrase accompanied with a subject particle, but the verbal phrase is always positioned at the end of sentence.

Hence, our annotation structure needs to be flexible enough to take into account the different ordering of words in sentences. In Figure 4, the nominal phrase *시스템을* (Engl. system) is an object and *공격하다* (Engl. attack) is a verbal headword. We can annotate the phrases according to their order in the sentence.

```
<participant type="Act" role="Destination">
 <syntactic-function name="Object" case="Accusative">
   <syntactic-type name="NP" particle="을">
     <realization > 시스템 </realization>
   을
   </syntactic-type>
 </syntactic-function>
</participant>
<lexie-att lemma="공격하다">공격하다가</lexie-att>
<lexie-att lemma="공격하다">공격하다가</lexie-att>다가
```

Figure 4: Position of the headword in the annotation

5.3 Syntactic function in Korean: annotation of cases

In Korean, we also encode case information as an attribute of the syntactic function tag as shown in Figure 4. In Figure 5, the nominal phrase plays subject role and the corresponding case is agentive. For the case feature, we can mark the realized value using case particles (lines 40 to 44 in Appendix B).

```
<syntactic-function name="Subject" case="Agentive"> …
</syntactic-function>
```

Figure 5: Case information in a syntactic function tag

5.4 Syntactic types

In Korean, adjectival phrases behave like verbal phrases, and prepositional phrases do not exist. We can thus find noun phrases, adverbial phrases[6] and clauses in our annotations. In the tags, we describe two features: the type of the phrase and the particle that expresses the case. Figure 6 shows how we encode the information (line 48, Appendix B).

```
<syntactic-type name="NP" particle="가"> …
 </syntactic-type>
```

Figure 6: Annotation of phrase type (SN)

[6] According to the particle classification of Chang (1996:56-61): "particles that mark oblique objects are called adverbial particles, for they function as adverbs, indicating location, direction, goal, source, and the like. (...) The combination of a noun and an adverbial particle forms an adverb phrase."

5.5 Prepositions in French

In French, when propositional phrases are used, a new attribute is added, namely *preposition*, and we indicate the preposition that appears in the context (e.g., *à*, *pour*, *de*). Figure 7 show how this information is taken into account in the annotations (line 47, Appendix B).

```
<lexie-att lemme="appuyer">appuie</lexie-att>
 <participant type="Act" role="Patient">
   <fonction-syntaxique nom="Complement">
       <groupe-syntaxique nom="SP" preposition="sur">sur
           <realisation>{Ctrl}-C</realisation>
       </groupe-syntaxique>
    </fonction-syntaxique>
</participant>
```

Figure 7: Adding an attribute for prepositions in French

5.6 List of cases and particles in Korean

While the syntactic function of phrases is chiefly reflected by their order in sentences in French, particles and cases indicate their syntactic function in Korean. Since Korean particles are very rich and developed, the list can be very long. However, in our annotations, particles are relatively limited because the contexts of verbal term entries are extracted from domain specific texts. The main cases and their corresponding particles found in terminological contexts are listed below.

Agentive case: 이/가 (Rom. -i/-ga)

Accusative case: 을/를 (Rom. –eul/-reul)

Dative case: 에/에게 (Rom. –e/-ege)

Locative case: 에/에서 (Rom. –e/-eseo)

Allative case: 로/으로 (Rom. –ro/-euro)

Instrumental case: 로/으로/으로써 (Rom. –ro/-euro/-eurosseo)

We added an attribute particle to our schema to account for this phenomenon (line 48, Appendix B).

Particles such as *은*, *는*, *도*, *만*, *조차*, *까지* are also used in contexts, but these are discourse function particles. Their behavior is different from other particles; they can be added to any other particle, even to verbs (공격하기도, 접속하기까지, etc.). Their function is to mark comparison, contrast, focus, limitation, etc.

5.7 Enumeration of terms instantiating an actant

In sentences, actants can be expressed by several terms. The enumerated terms for one actant are annotated by repeating the participant tag.

(5) *두 대의 컴퓨터가 하나의 모니터 및 마우스와 스피커까지 공유할 수 있다* (Engl. Two computers can share a monitor, a mouse, and even a speaker.)

In sentence (5), we can see that three terms *모니터* (Engl. monitor), *마우스* (Engl. mouse), *스피커* (Engl. speaker) instantiate the PATIENT role. However, in such as enumeration in Korean, only the last term will carry a particle indicating the accusative case. Since we list participants in enumerations separately, our annotation will mention the particle for each different participant.

In addition, terms in example (5) are used with comitative case particles (which in this example indicate coordination). If comitative cases are seen in enumerations, we do not annotate them.

5.8 Relationships between elements of a phrase

In Korean specialized texts, the chain <noun + generic case particle + noun> is often found. It corresponds to the French pattern <noun + de + noun>.

(6a) *네트워크 설정* (Engl. configuration network)

(6b) *시스템 지원* (Engl. system configuration)

In noun phrases such as those in (6), modifiers carry particles expressing genitive cases. These are not taken into account when annotating verbs, since only heads of syntactic groups are analyzed. However, they would be taken into account in the annotation of nouns.

6. Conclusion and future work

As far as semantic roles are concerned, it appears that our annotation model, previously designed for French contexts, can be extended to Korean. Our work on samples, i.e. on verbal headwords *공격하다* (Engl. attack), *공유하다* (Engl. share), *개발하다* (Engl. develop), *검색하다* (Engl. search), *구현하다* (Engl. implement), *나누다* (Engl. partition), *변환하다* (Engl. convert), *삭제하다* (Engl. delete), enabled us to verify that the model can be applied to Korean. Considering that these two languages are very different, we are led to believe that the original semantic annotation model can be used for other languages.

Of course, some adaptations are required to take into account the syntactic characteristics of each language. For instance, the preposition tag is no longer necessary in Korean; on the other hand, we added two features to take into account particles and cases

Regarding these latter specificities, it would be interesting to list specific syntactic phenomena linked to the expression of semantic roles.

In Korean annotations, we could inventory which semantic roles accompany which cases and particles, and confirm the constraints of particle usage in domain-specific texts. Compared with general texts in which style is important, domain-specific texts use a restricted set of particles. However, since this work is only at an early stage, an optimized list of semantic roles and an inventory of Korean cases for each actant and the semantically annotated contexts will be developed continuously.

Similarly, the annotation of French contexts will enable us to list syntactic functions, types and

prepositions that can be linked to specific semantic roles. While being aware that functions, types and prepositions can be ambiguous (i.e., correspond to more than one semantic role), these lists could be useful to recognize roles in French sentences.

Hence, the relationship between the syntactic and the semantic structures will be investigated further. For instance, in Korean, a particle can be embodied in several cases in various contexts, and a case can present several semantic actants. For example, 로 is a particle for marking the indirect complement and that represents the Directional case. In addition, this case can represent the semantic roles corresponding the DESTINATION or DESTINATAIRE. However, this particle can be used to express very different cases: instrumental, transformative, directional, etc. from one contact to another. In addition, these cases should be subcategorized (Cho & Kim 1995).

Instrumental :

도구/수단/재료 Instrument/Means/Material

Transformative:

변성/분할/선정 Change/Division/Selection

Directional:

경로/방향/지향점 Path/Direction/Destination

Acknowledgments

This research was supported by the Social and Humanities Research Council of Canada (SSHRC). The authors would like to thank Fadila Hadouche who has participated in the design of the XML schema and Stéphanie Caron, Stéphanie Klebetsanis, Annaïch Le Serrec and Charlotte Tellier who have annotated the French contexts.

References

L'Homme, M.C., Bae, H.S. 2006. A Methodology for Developing Multilingual Resources for Terminology. In *LREC 2006. Language Resources and Evaluation. Proceedings.*

Baker, C. F. and J. Ruppenhofer. 2002. FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (eds.) *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society. 27-38.*

Baker, C.F. C.J. Fillmore and J.B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Chang S.J. 1996. *Korean*, John Benjamins, Amsterdam/Philadelphia.

Cho I.Y. and Kim I.H. 1995. Case particle –로(-Ro) and its semantic role (격조사 –로의 의미역). In *Korean linguistics (국어학)*. CRBook. pp.1-22.

Dolbey A, Ellsworth, M., Scheffczyk, J. 2006. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In O. Bodenreider (ed.). *Proceedings of KR-MED*, 87-94.

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

Fillmore, C.J. 1968. The case for case, In Bach, E. and R.T. Harms (eds.). *Universals in Linguistic Theory*, New York: Holt, Rinehard and Winston, 1-88.

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

Kicktionary. (http://www.kicktionary.de/) Accessed 10 September 2006.

Lee K.J. 1999. Research on case in Korean traditional grammar (전통문법에서의 격연구). In *Case and particle in Korean (국어의 격과 조사)*. Association of Hangeul, Worin, Seoul. pp.9-48.

Ohara, K. H., Fujii S., Ohori T., Suzuki R., Saito H., Ishizaki S. 2006. Frame-based Contrastive Lexical Semantics and Japanese FrameNet: The case of RISK and *kakeru*. In *ICCG4 (the Fourth International Conference on Construction Grammar)*, Tokyo, Japan.

Ohara, K. H., Fujii S., Ohori T., Suzuki R., Saito H., Ishizaki S. 2004. The Japanese FrameNet Project: An introduction. LREC 2004. The Fourth international conference on Language Resources and Evaluation. *Proceedings of the Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora"*, 9-11.

Ruppenhofer, J., M. Ellsworth, R.L.M. Petruck, C. Johnson and J. Scheffczyk. 2002. *FrameNet II: Extended Theory and Practice.*

Schmidt, T. forthcoming. *The Kicktionary – A Multilingual Lexical Resources of Football Language*. (http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126). Accessed 12 September 2008.

Petruck, M.R.L. 1996. Frame Semantics. In J.Verschueren. J-O. Oestman, J. Blommaert, and C. Bulcaen (eds.). *Handbook of Pragmatics*. Philadelphia: John Benjamins.

You L., Liu T., Liu K. 2007. Chinese FrameNet Data in Semantic Web Language. In *Proceedings of the Conference on Natural Language Processing and Knowledge Engineering*. 50-55.

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

Appendix A: A sample of French annotated contexts and the HTML interface to validate them

```
<vocable identificateur="configurer">
      <lexie numero-acception="1">
    <contexte source="CODERREUR" statut="1" annotateur="SK MCLH" mise-a-jour="2008-03-06"> [25 lines]

    <contexte source="BIOSRV" statut="1" annotateur="SK" mise-a-jour="2008-02-23"> [25 lines]

    <contexte source="CEVEIL" statut="1" annotateur="SK" mise-a-jour="2008-02-23">
      <contexte-texte>Ainsi, lorsque l'usager configurerait son nouveau micro-ordinateur, il indiquerait quels sont les paramètres de sa localisation physique
        et de sa culture.</contexte-texte>
      Ainsi, lorsque
      <participant type="Act" role="Agent">
        <fonction-syntaxique nom="Sujet">
          <groupe-syntaxique nom="SN"> l'
            <realisation>usager</realisation>
          </groupe-syntaxique>
        </fonction-syntaxique>
      </participant>
      <lexie-att lemme="configurer">configurerait</lexie-att>
      <participant type="Act" role="Patient">
        <fonction-syntaxique nom="Objet">
          <groupe-syntaxique nom="SN">son nouveau
            <realisation>micro-ordinateur</realisation>
          </groupe-syntaxique>
        </fonction-syntaxique>
      </participant>, il indiquerait quels sont les paramètres de sa localisation physique et de sa culture.
    </contexte>

    <contexte source="CODERREUR" statut="1" annotateur="SK" mise-a-jour="2008-02-23"> [11 lines]
  </lexie>
</vocable>
```

CONFIGURER 1

Si ***vous*** voulez installer ou **CONFIGURER** **un** ***logiciel*** sur le *serveur* , contactez votre administrateur réseau. [CODERREUR 1 SK MCLH 2008-03-06]

En **CONFIGURANT** *correctement* **la** ***mémoire cache*** , ***on*** peut améliorer considérablement les performances de l'ordinateur. [BIOSRV 1 SK 2008-02-23]

Ainsi, lorsque **l'** ***usager*** **CONFIGURERAIT** **son nouveau** ***micro-ordinateur*** , il indiquerait quels sont les paramètres de sa localisation physique et de sa culture. [CEVEIL 1 SK 2008-02-23]

Pour **CONFIGURER** ou supprimer **la** ***version*** **existante de ce produit** utilisez Ajout Suppression de programmes depuis le Panneau de configuration. [CODERREUR 1 SK 2008-02-23]

CONFIGURER 1		
Actants		
Agent	Sujet (SN) (2) Lien indirect (SN) (1)	on usager vous
Patient	Objet (SN) (4)	logiciel micro-ordinateur mémoire cache version
Autres		
Lieu	Complement (SP -sur) (1)	serveur
Manière	Modificateur (SAdv) (1)	correctement

Appendix B: Schema used for the annotations (The Schema uses the RelaxNG compact notation).

```
start = element vocables{element-vocable*}

element-vocable = element vocable {
    attribute identifier {text},
    element-lexie*
}
element-lexie = element lexie {
    attribute sense-number {text},
    element-context*
}
element-context = element context {
    attribute source {text},
    attribute state {xsd:nonNegativeInteger},
    attribute annotator {list{TypeAnnotator*}},
    attribute update {xsd:date}?,
    element-context-text,
    mixed {(element-lexie-att | element-participant | element-antecedent)*}
}
element-context-text = element context-text {text}

element-lexie-att = element lexie-att {
   attribute auxiliary{Auxiliary}?,
   attribute lemma {text}?,
    text
}
element-participant = element participant {
   attribute type {TypeParticipant},
   attribute role {RoleParticipant},
   element-syntactic-function
}
element-antecedent = element antecedent {
   attribute xml:id {xsd:ID},
   mixed {element-value-antecedent*},
   text
}
element-value-antecedent = element value-antecedent{
attribute lemma{text}?,
text
}
element-syntactic-function = element syntactic-function {
   attribute name {NameSyntacticFunction},
   attribute case {CaseSyntacticFunction}?,
   element-syntactic-type
}
element- syntactic-type = element syntactic-type {
   attribute nom {NameSyntacticType},
   attribute preposition {text}?,
   attribute particle {text}?,
   mixed {element-realization}
}
element-realization = element realization{
   attribute lemma {text}?,
   attribute semantic-label {text}?,
   attribute ref {xsd:IDREF}?,
   text
}
```

Multilingual modalities for specialised languages

Yukie Nakao
LINA, Nantes University, 2, rue de la Houssinière, BP 92208 44322 Nantes Cedex 03, France
Yukie.Nakao@univ-nantes.fr

Abstract

Specialised language is the language used in a specific domain by specialists. It is derived from general language, but the concrete differences between a given specialised language and a general one must be realised in terms of linguistic features. This paper focuses on linguistic features involving modality, and whether the presence or absence of these features can reveal differences between different discourse types in specialised languages. We consider how a theory of modality can be applied to a language which realises that modality in a different way, and compare the application of two different modality theories, locutionary modality and irrealis modality, in a French-Japanese medical corpus. Primarily due to its semantic character, the locutionary modality theory appears to define a clearer relationship between characteristics of utterances and modalities.

Introduction

Specialised language is by definition the language used in a specific domain by specialists. Specialised languages are derived from general ones, but restrict them in various ways (Sager *et al*, 1980). However, the concrete ways in which a given specialised language differs from a general one must be realised in terms of the linguistic features present in the general language, which will not necessarily be shared with other languages. Our particular interest here will be to focus on linguistic features involving modality, and whether the presence or absence of these features can reveal differences between different discourse types in specialised languages. In general, an utterance can convey two different types of information: propositional content, and the speaker's (writer's) relationship to it. The relationship between speaker and propositional content, sometimes called *modality*, declares the speaker's subjectivity, his relationship with his addressee, and so on. For example, in the sentence "she believes that they went away", the *proposition* of the utterance is "they went away". The *modal* marker "she believes" declares what the speaker's attitude is to this proposition. In European languages, modality is most often realised using verbal constructions. This is by no means universally the case, however, and modality can often be a key item to consider when comparing the structures of different languages. The central question we will consider in the paper is: how can a theory of modality, initially defined in terms of the concrete linguistic structures of one language, be applied to a language which realises that modality in a radically different way? We will investigate the issues concretely, by comparing the application of two different modality theories, locutionary modality and irrealis modality in a French-Japanese medical corpus. In particular, we will analyse how specific linguistic features are involved in the application of these modality theories.

1. Modality

As already indicated, modality essentially expresses the speaker's opinions or attitudes towards the core content of the utterance (Palmer, 1986, Le Querler, 1996, Miyazaki *et al.*, 2002). Palmer points out the typology of modality helps to identify the grammatical categories between different languages (*ibid.*). This remark is very much in line with our goal in this paper, namely the analysis of a multilingual corpus. This section presents two modality theories, namely the locutionary modality and the irrealis modality.

1-1. Locutionary modality

Locutionary modality examines the speaker's functions in the utterance. In the interests of using a framework that is relatively language-neutral, our analysis here will follow the work of Charaudeau (1992). Written for French, this theory is primarily formulated in terms of verbal constructions. However, since the analysis of these constructions is based on semantic universals, application to Japanese seemed feasible. In this context, we also consulted some studies of the characteristics of Japanese modality (Miyazaki *et al*, *ibid*., Masuoka, 1992). The principal notion of modality as formulated by Charaudeau is defined by the three basic locutionary acts: the Addressee's act, the Speaker's act and the Illocutive act. In the Addressee's act, the speaker implies his addressee in the utterance and imposes on the latter the content of the proposition (ex. *You should take an early train*). This act is classified into nine categories such as "injunction", "permission", "judgement", "suggestion" and so on. In the Speaker's act, the speaker makes no direct reference to his addressee (*Because of the strike, I had to walk*). This act is classified into twelve modalities, such as "constatation", "knowledge/lack of knowledge", "opinion", "obligation" and so on. Finally, the Illocutive modalities refer neither to the speaker nor to the addressee (ex. *It is clear that she is wrong*). Our object in using this theory is to determine the function of the locutionary functions in multilingual discourse analysis. The Illocutive modalities are not treated here.

1-2. Irrealis modality

Another modality theory, *irrelis modality* (Givón, 1994, Nomura, 2003) examines whether the proposition in an

utterance is realised or not. When the speaker believes that the event has been realised, the utterance is categorised as *realis*. When the speaker believes that the event has not been realised, the utterance is categorised as *irrealis*. As Takahashi (2007) remarks, for the languages that distinguish *realis/irrealis* moods by their morphologic representations such as inflection, these moods are mainly predicate features. For example, Givón presents the grammatical features for English as: future tense (*I will come tomorrow morning*), modal adverbs (*maybe she is right*), command (*go away!*), exhortation (*let's go home*), yes-no question (*have you read this article?*), subjunctive (*I shall talk to her when she arrives*) and modal auxiliaries (*they must study more*). For French, we need to add the verb forms *conditionnel*, *imparfait* and *futur antérieur* as markers of irrealis. In the case of Japanese, which belongs to the agglutinative languages, Tamachi (2005) applies Palmer's theory (2001) and classifies the modalities by verbal forms, compound forms, auxiliary verbs and modal equivalents.

2. Corpus

Our objective in using these modalities is to apply them to a multilingual corpus for a specialised language. This section describes the structure of the corpus and the modality criteria to be examined.

2-1. Comparable corpus

We applied the above mentioned modalities within the scope of a comparable corpus for French and Japanese. The comparison is carried out along two separate dimensions (Zweigenbaum, 2006). The first dimension is topic. Here, we target a common topic shared by the documents of the two languages; we chose the medical domain and the specific topic of "diabetes and nutrition". The second dimension is that of distinctions of communicative level, i.e. discourse types. Discourse types are considered to be the most general level of textual classification by Malrieu and Rastier (2002). We distinguish between scientific articles proper (by definition, written for specialists), and popular science discourse (by definition, written for non-specialists). The documents written for specialists turned out to have a richer vocabulary and more complex content. Based on these ideas, our corpus consisted of a collection of Web documents for the two languages.

2-2. Modality criteria

For the application of two modality theories, we listed the criteria for each category. For the locutionary modality, the modality markers are mainly verbs. For example, the modality of suggestion is expressed with the verbs like *suggérer* (suggest) *conseiller* (advise) for French, 勧める (recommend) for Japanese. Use of the *–ons* form (French) or the -ましょう form (Japanese) in the main verb (similar to English "let's…!") are also treated as markers expressing suggestion. In the case of the irrealis modality, inflected forms of verbs and modal auxiliaries are used as criteria in French, and verbal forms in Japanese. For example, the conditionnel of French has the conjugation of *–rais, -rait, -riez, -rions* and *–raient*. For Japanese, the verbal form like かも知れない (might) indicates the lower certainty of the fact. We calculated manually the number of occurrences labelled according to each modality theory.

For this process, we made a test corpus constituted of five documents derived from the above mentioned corpus for each discourse type in each of the two languages. The minimum, maximum and average number of the utterances in the test corpus for each discourse type are as follows:

French
-Scientific discourse: 16-63 (av. 30.4)
-Popular scientific discourse: 11-63 (av: 34.6)
Japanese
-Scientific discourse: 33-49 (av. 49.4)
-Popular scientific discourse: 13-98 (av. 44.8)

Table 1 and Table 2 show, for each modality, the number of occurrences of modal markers found in the test corpus. The number in parentheses indicates the number of cases where the markers were unclear. When an utterance contained more than one modal marker, we calculated all the possibilities. Some examples from the test corpus are also shown in Table 3 and Table 4 at the end of this paper.

3. Contrastive analysis of two modality theories

In our corpus, the texts normally have a neutral nature, and the speaker does not appear so often. The results show that the irrealis modality always has more markers than the locutionary modality, which is based on the speaker's functions in the utterance. The two modality theories also reveal the similarity of the results for the two discourse types in French. In Japanese, scientific discourse has less modal markers than popular scientific discourse for both modalities.

3-1. Locutionary modality

The result of the locutionary modality proposes various characteristics. As a whole, the modalities of interrogation and of opinion are common categories for both discourse types and for both languages. The modalities of estimation and of suggestion are mainly found in the Japanese corpus. The number of occurrence of the modality of suggestion is especially high in Japanese popular scientific texts. This reflects the pedagogic purpose of the Japanese popular scientific texts, which often explain dietary restrictions.
In French, a list of verb phrases in the infinitive form is often used in place of a construction which is explicitly marked to indicate the modality of suggestion. This explains the absence of this modality in French corpus. In the French corpus, appearances of modal markers are

not as frequent on the whole. The French texts tend to be descriptive and the utterances are typically expressed using impersonal form. If this is a general tendency of the whole French corpus, this might be interpreted as suggesting that the text has a neutral nature. The corpus used is in the medical domain, and most of the texts are descriptive narratives written in the first person. The way in which modality is realised is influenced by the discourse type.

On the other hand, as the many numbers in parenthesis show, the weak point of the locutionary modality theory is the difficulty of determining the locutionary markers. For French, the impersonal pronoun "on" is frequently used. "On" has four different meanings:1) *someone* 2) *people* 3) *everyone* 4) *we*. What we are interested is "*on*" as *we*, but it is not easy to distinguish what "*on*" in the corpus means.

	fr		**jp**	
	sc	*po*	*sc*	*po*
ADDRESSEE'S ACT				
Call	-	-	-	-
Injunction	-	-	-	4(1)
Permission	-	-	-	-
Warning	-	-	-	2
Judgement	1	-	-	-
Suggestion	1	-	1(1)	43
Proposition	-	-	-	-
Interrogation	12	7	1	6
Request	-	-	-	-
SPEAKER'S ACT				
Constatation	-	2(1)	2	-
Knowledge/Lack of knowledge	-	-	-	-
Opinion	2	1(1)	7(7)	1
Estimation	-	-	4	2
Obligation	-	-	-	-
Possibility	-	5(5)	-	-
Desirability	-	-	1(1)	-
Promise	-	-	-	-
Acceptation/refusal	-	-	-	-
Agreement/disagreement	-	-	-	-
Declaration	-	-	-	-
Proclamation	-	-	-	-
TOTAL	**16**	**15(7)**	**16(9)**	**58(1)**

Table 1. Number of occurrence of the locutionary modality

For Japanese, ambiguity is caused because of the phenomenon of zero anaphora. As is well known, this is a general phenomena and can occur in any genre of discourse. When this happens in an utterance, the locutionary information is determined by context or the politeness markers indicated by particles and special lexical forms. However, the difficulty remains when the first person and second person are both possible. From these points of view, our hypothesis is that these ambiguities are not concerned with other modality theories that do not concern the speaker's function in the utterance.

3-2. Irrealis modality

Moving on to the irrealis modality, for Japanese, exhortation and adverbial markers show up a difference between two discourse types. In general, there are more utterances in the popular scientific discourse with these markers. For French, the total counts are similar between the two discourse types. No markers were found for the modality of modal adverbs, command, and jussive. As for the modality of exhortation, no utterance was found in scientific discourse. Modality markers for non-declarative speech acts were not frequent in the French corpus.

		fr		**jp**	
		sc	*po*	*sc*	*po*
Future, futur antérieur		4	3		
Comlements of modality verbs		8	6	5(2)	4
Non-declarative speech-acts	Command	-	-	-	4(1)
	Request	-	-	-	-
	Exhortation	1	-	1	43
	Jussive	-	-	-	-
	YesNo questions	5	2	-	1
Adverbial clauses		4	7	2	24(3)
Modal auxiliaries		7	14	12	10(4)
Conditionnels		5	7		
TOTAL		**36**	**39**	**20(2)**	**86(8)**

Table 2. Number of occurrence of the irrealis modality

With the irrealis modality, the difficulty is to determine whether an utterance is irrealis or not. For example, in the utterance 放っておくと糖尿病に移行する可能性が高いのが境界型 (the type that risks acquiring diabetes is the pre-diabetic type), the modal marker is in the subordinate clause, but the whole utterance can be irrealis or realis, since it talks about general knowledge of diabetes. Similar problems are found in the French corpus. In the utterance "les comprimés permettent schématiquement d'agir à trois niveau" (the pills can schematically act at three levels), we judged that the verb "permettre" was a higher verb marking modality. However, this verb can be realis or irrealis. Givón (1994) classifies this sort of habitual modality marker as hybrid.

4. Conclusion and Future work

The tendencies seen in the analyses according to the two modality theories are not identical, because of their difference in approach. The locutionary modality theory makes the relationship between characteristics of utterances and modalities clearer. The fact that the definitions are primarily semantic in nature (even though they also make reference to grammatical categories) doesn't limit them to a single language. Their application to Japanese was in fact proved to be quite feasible. Concerning the irrealis modality, the differences in tendencies between the two languages are obvious given their grammatical characteristics. Also another advantage is the possibility of less arbitrary judgement when

determining each modal marker, since this modality is primarily based on grammatical categories. As future work, we are considering investigating the roles of the modalities in the scope of specialised language. Also we hope to find the modalities' characteristics from the point of view of the relationship between discourse types and terminology.

5. References

Charaudeau, P. (1992). *Grammaire du sens et de l'expression*, Hachette.

Givón, T (1994). "Irrealis subjunctive", *Studies in language*, 18 (2), p. 265-337.

Le Querler, N. (1996). *Typologie des modalités*, Presses universitaires de Caen.

Malrieu, D. & Rastier F. (2002). "Genres et variations morphosyntaxiques", *TAL*, 42 (2), p. 548-577.

Masuoka, T. & Takubo Y. (1992). *Basic Japanese grammar, Kuroshio-syuppan* (Japanese)

Miyazaki, K., Adachi, T., Noda, H. & Takanasi, S. (2002). *Modality*, Kuroshio-syuppan (Japanese).

Nomura, T. (2003). "Typology of the modality form", *Kokugogaku*, The Society of Japanese Linguistics, 54 (1), p. 17-31 (Japanese).

Palmer, F. R. (1986). *Mood and Modality*, Cambridge University Press.

Palmer, F. R. (2001). *Mood and Modality*, Cambridge University Press. 2nd edition.

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

Takahashi, K. (2007). "Irrealis modality markers in Thai", *Kanda Gaigo Daigaku Kiyou*, 19, p. 1-22.

Tamachi, M. (2005), "A contrastive study of Modality in Japanese and Chinese -A Linguistic Typological Perspective-", *Takamatsu Daigaku Kiyou*, no.14, p. 17-54.

Zweigenbaum, P. & Habert, B. (2006). "Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue", *Glottopol*, 8, p. 22-44.

1	(…)必ず医師の指示に従ってください。 (**Follow** (your) doctor's opinion) Imperative, JP-*po*
2	**Revenons** donc aux malheureux engagés à leur insu sur la voie du diabète. (**Let's go back** to the unhappy people who are unknowingly on the road to the diabetes) Suggestion, FR-*sc*
3	(…) **on observe** dans de nombreux pays (…) une augmentation particulièrement importante du diabète de type 2 (…). (We observe in many countries a significant increase in Type 2 diabetes) Constatation, FR-*po*
4	(…) たいへん意義深いものであると思われる。 (**(We) think** it quite meaningful) Opinion, JP-*sc*
5	本特集が、この分野の研究の理解・発展の一助となれば私とすれば**このうえない幸いである**。 (**I will be grateful** if this special issue adds to the comprehension and development of research in the domain) Estimation, JP-*sc*

Table 3. Examples of the locutionary modalities, taken from the test corpus. Each example is accompanied with its English translation as well as the modality criteria. JP: Japanese, FR: French, *po*: utterances of popular scientific discourse, *sc*: utterances of scientific discourse.

6	le conseil nutritionnel **pourra** aller vers des aliments plutôt riches en potentiel anti-oxydants. (Advice about food **will** primarily be directed towards foods rich in potential anti-oxidants) Future, FR-*sc*
7	塩分を制限されている方は注意して！ (*Be careful* if you are on a low-sodium diet!) Non-declarative speech acts/command, JP-*po*
8	**Quand** le poids monte, les glycémies montent. (**When** your weight increases, glycerine levels also increase) Adverbial clause, FR-*sc*
9	La part des graisses **doit** être réduite même si la tendance spontanée **serait** plutôt de l'augmenter en réaction à la limitation des apports en glucides. (Fat content **must** be reduced, even if the spontaneous tendency **would** rather be to increase it in reaction to the reduced contribution of the glucose) Modal auxiliary and conditionnel, FR-*po*

Table 4. Examples of the irrealis modalities, taken from the test corpus. Each example is accompanied with its English translation as well as the modality criteria. JP: Japanese, FR: French, *po*: utterances of popular scientific discourse, *sc*: utterances of scientific discourse.

A Categorisation Framework Editor for Constructing Ontologically underpinned Terminological Resources

Peter De Baer1, Koen Kerremans2, Rita Temmerman3
Erasmushogeschool Brussel – Centrum voor Vaktaal en Communicatie
Trierstraat 84, 1040 Brussels, Belgium
E-mail: peter.de.baer1@ehb.be, koen.kerremans2@ehb.be, rita.temmerman3@ehb.be

Abstract

In this article we describe the Categorisation Framework Editor (CFE), a software tool to construct ontologically underpinned terminological resources. The CFE supports the Termontography-methodology which is a multidisplinary approach in which theories and methods for multilingual terminological analysis of sociocognitive theory (Temmerman, 2000) are combined with methods and guidelines for ontology engineering. A clear distinction is made between conceptual modelling at a language-independent level and a language-specific analysis of units of understanding. The data structure used to store the terminological information we call a Categorisation Framework and this structure will also be described in this article. Since the CFE was developed during the PoCeHRMOM-project, we will illustrate the use of the CFE with examples from this project. During the PoCeHRMOM-project an online platform, named Profile Compiler, was developed on which companies and organisations could create job profiles. A job profile typically contains a list of required competencies and qualifications, together with a list of tasks a job candidate should be able to carry out. The CFE was used to develop the ontologically structured multilingual (EN, FR, NL) terminological database for the project.

1. Introduction

In this article we describe the Categorisation Framework Editor (CFE), a software tool to construct ontologically underpinned terminological resources. The CFE supports the Termontography-methodology which is a multidisplinary approach in which theories and methods for multilingual terminological analysis of sociocognitive theory (Temmerman, 2000) are combined with methods and guidelines for ontology engineering. A clear distinction is made between conceptual modelling at a language-independent level and a language-specific analysis of units of understanding. In Termontography (Temmerman & Kerremans, 2003) a knowledge analysis phase should ideally precede the methodological steps which are generally conceived as the starting-points in terminography: i.e. the compilation of a domain-specific corpus of texts and the understanding and analysis of the categories that occur in a certain domain. This view results from the fact that terminological databases need to represent in natural language those items of knowledge or 'units of understanding' which are considered relevant for specific purposes, applications or groups of users. In Termontography, the units of understanding as well as their intercategorial relations are therefore structured in a categorisation framework (Kerremans, 2004). On the one hand, this framework supports the information gathering phase during which a corpus is developed. On the other hand, it allows terminographers to establish specific extraction criteria as to what should be considered a 'term': i.e. the natural language representation of a unit of understanding, considered relevant to given purposes, applications or groups of users.

The CFE was developed during the PoCeHRMOM-project[1] therefore we will illustrate its use with examples of this project. During the PoCeHRMOM-project an online platform, named Profile Compiler, was developed on which companies and organisations could create job profiles. This project and the role of the CFE therein will be described in part 4 of this article. In part 2 we will first describe the data structure we use to store both the terminological and the ontological information. In part 3, we will discuss the implementation of the CFE and its advantages for terminography and linguistic ontology engineering. In part 5 we describe some related and possible future work and in part 6 we formulate a conclusion.

2. The Categorisation Framework

The CF is designed to represent linguistic ontologies, i.e. to describe ontologies merely using terminology. For this purpose the CF consists of the following 8 items: language string, category, meta category, term, meaning, bi-directional relation, bi-directional relation instance, and property.

A **language string** is a character string that is classified by one or more categories. Each language string must be classified by just one language category.

For instance, the character string “main” classified by the French language category “fr”[2] constitutes the French language string “main”, i.e. hand.

A **category** belongs to a single context we call a meta category. A category can be used to classify CF items, for instance language strings.

The English language category “en”, for instance, belongs to the meta category “language”.

A **meta category** specifies the context of all the categories that belong to it.

If we add the category “fr”, for instance, to the meta category “language”, this category represents the French language.

[1] A Flanders IWT-TETRA funded project to support e-HRM for SMEs using Semantic Web technology.

[2] Following ISO 639-1: codes for the identification of languages.

By means of these three CF items meta category, category and language string we are able to contextualize terminology. For example, we could create a meta category "regional language" and add the category "en-UK"[3] to this meta category. This category "en-UK" could then be used to classify the English language string "aubergine" as a British English word (the common American English word for this fruit is "eggplant").
Although we now have a simple classification mechanism to add context to terminology, there are still some issues that must be resolved. Our goal is to use terminology to specify concepts and concept relations. We can see that a character string is insufficient to refer to a certain meaning. The character string "main", for instance, might refer to a "hand" when interpreted in French or a "chief or largest part" when interpreted in English. For this reason we introduced the CF item language string that is a combination of a language category and a character string. This does not suffice, however, to solve the problems of polysemy and semantic vagueness. The English language string "bow", for instance, might refer to "the front part of a ship", "to bend", "a weapon", etc. To resolve this semantic ambiguity, we provide context to the language strings in order to disambiguate them. For instance, if we want to indicate the concept "weapon that shoots arrows", we might relate the hypernym "weapon" with the English language string "bow". Meta categories can be used to provide this necessary context. The previous three meanings of the English word "bow" could be implemented as described in table 1.

Meta category	**Category**	**Category description**
part of a ship	bow	front part of a ship
verb	bow	to bend
weapon	bow	weapon that shoots arrows

Table 1: Three meanings of the English word "bow".

We should note that the identification of meta categories and categories, as is shown in table 1, is no longer by means of language strings. Instead, we use the CF item term that refers to a language string in combination with a specific meaning.
A **term** is an item with both a reference to a language string and a meaning, i.e. (meta) category. Several terms may refer to a single meaning.
With this definition a term now refers to a certain meaning. The term that references the English language string "bow" and the category "bow" with meta category "weapon", for instance, clearly describes this term as a weapon that shoots arrows.
We should also note that a meta category might be a category itself. Consider for instance the meta category "verb", this meta category could be a category itself with meta category "part of speech". To practically implement this, we further extend the CF with a meaning item.
A **meaning** item is the underlying CF item of a meta category and/or a category. A meaning item has a list of terms and may have references to both a meta category and a category.
The term "verb", for instance, could reference a certain meaning item with references to the meta category "verb" and the category "verb".
By linking terms to a (meta) category using the underlying meaning item, the meaning of the (meta) category will be specified.

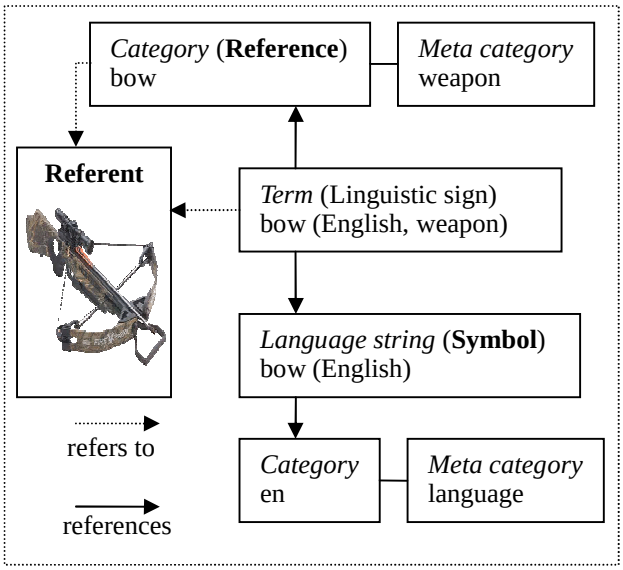


Figure 1: The CF items (*italics*) in relation to the semiotic notions (*bold*).

Figure 1 summarizes the CF items defined in this section and compares them with better known notions from semiotics (Chandler, 2006). The two categories and the two meta categories should also be considered as meaning items. The symbol is the character string "bow" that is classified by the English language category "en" and hence constitutes an English language string. The term that references both the English language string "bow" and the category "weapon that shoots arrows" now clearly represents a specific meaning. This meaning directly refers to the intended referent. Consequently, the term indirectly refers to the intended referent.
By now we are able to represent concepts, by means of terminology (using terms). In a similar way we can identify concept relations if we add the meta category "relation" and add for each concept relation a specific category to this meta category. For instance the category "has meronym" could represent the whole-to-part relation. To create relations between meanings and/or terms, we include the CF items bidirectional relation and bi-directional relation instance.
A **bi-directional relation** references at least one forward

[3] Following RFC 3066: codes for the identification of (regional) languages.

relation and possible also a backward relation.
A relation can be created by adding a category with meta category "relation". To the meta category "relation", for example, we may add the categories "has meronym" and "has holonym". A logical bi-directional relation could reference both opposing categories. Let us notate this bi-directional relation as the ordered pair [has meronym; has holonym].
A **bi-directional relation instance** references a bi-directional relation and two meaning items or two term items. Since the direction of a bi-directional relation is usually relevant, the bi-directional relation instance makes a distinction between the source and the target item.
The bi-directional relation instance that references the bi-directional relation [has meronym; has holonym], the source category "arm" and the target category "lower arm" would, for instance, indicate that a "lower arm" is part of an "arm".
Using bi-directional relation instances, the CF can be structured in a flexible, yet generic manner. A taxonomy could be created, for example, based on the list of meta categories. By doing so, this meta category hierarchy may be used to browse the CF. Since each meta category may have multiple parents, multiple entry points can be provided to expose the underlying categories. Besides providing context to categories, the hierarchy of meta categories would thus have an extra function comparable to that of topics in a topic map[4].
The hierarchical generic-specific relation could be implemented by means of the bidirectional relation [has hyponym; has hypernym] between categories. By adding relations and bi-directional relations all types of concept relations can be used to build a concept model. For example, the bi-directional relation [requires; is required for] between the occupation "architect" and the competency "technical planning" indicates that an architect must have the competency "technical planning".
Similarly, a terminological network may be created by adding bi-directional relation instances between terms. The bi-directional relation [has lemma; is lemma of] between the terms "architects" and "architect", for instance, indicates that the lemma of "architects" is "architect".
The last CF item, i.e. property, we have to define may be used to provide extra information about a CF item.
A **property** references an attribute and a value of a certain value type. Properties may be added to each CF item i.e. language string, category, meta category, term, meaning, bidirectional relation, bi-directional relation instance, and property.
An attribute is implemented as a category with meta category "attribute". Each attribute should refer to a certain value type. The list of possible value types depends upon the specific implementation of the CF. The value type "character string" is a minimum requirement. The value types "URI"[5], "URN"[6] and "URL"[7] have been proven very useful too. For instance, the attribute "description" with value type "character string" could be used to describe a term, while the attribute "extra information" with value type "URL" could be used to refer to a web page with extra information about a category. A property with attribute "extra information" and value "http://en.wikipedia.org/wiki/Architect" could, for example, be linked to the category "architect" with meta category "occupation".
These examples show that the CF is a flexible and expandable structure to handle terminological and ontological information.

3. The Categorisation Framework Editor

The CFE is a Java application which makes it platform independent. To handle multilingual information the software uses the Unicode UTF-8 format. A CF application programming interface is used as the basis for the application in order to be able to exchange the CF information with other software applications. Currently the CFE is able to store a Categorisation Framework (CF) in CF XML format and CF JavaDB format. CF XML makes it possible to convert existing structured information resources into a CF, using simple scripts (e.g. Per[8]). This feature may facilitate the integration of existing terminological resources.
The applied CF structure makes it possible to merge multilingual and heterogeneous CFs based on common terminology. This is possible since each concept is identified by at least two terms, one term for the category and the other term for the meta category. The first term represents the concept itself, and the second term represents the context of the concept. This context might be for instance the hypernym or the holonym of the concept. For example, to represent the occupation "engineer" we might use the terms "engineer" and "occupation". The first term represents, what we call the category, and the second term represents, what we call the meta category. As stated above, the meta category "occupation" specifies the context, in this case the hypernym, of the category "engineer". The advantage of this approach is that we can uniquely identify a concept, merely by using terminology. The term "engineer" could still be used to represent a different concept, for example a qualification, but then the concept should have a different context. In this case, the category "engineer" could have, for instance, the meta category "qualification". Both a meta category and a category may be represented using

[4] Topic maps are an ISO standard for the representation and interchange of knowledge, with an emphasis on how retrievable information is. The standard is formally known as ISO/IEC 13250:2003.

[5] Uniform Resource Identifier

[6] Uniform Resource Name

[7] Uniform Resource Locator

[8] Perl (Practical Extraction and Report Language) is a dynamic programming language created by Larry Wall.

several (multilingual) terms. For example, the category “engineer” might also be represented by the German term “Ingenieur” and the meta category “occupation” might also be represented by the German term “Beruf”. Using the overlapping terminology it is then possible to automatically merge different CFs.
In contrast to the use of sets of synonymous words (synsets) in the EuroWordNet format (Vossen, 1998), the CF makes use of meaning items to bundle both synonyms and translation equivalents. The latter approach doesn't require the construction of an artificial Inter-Lingual-Index and facilitates the development of multilingual terminological resources that include aspects of cultural variation. Classifying categories and properties assigned to terms may be used for denotation of cultural variation. We could, for example, describe the legal differences of the concept “maison de repos” in Belgium and France. In both legal systems the same term may be used to refer to a concept with a similar general meaning, however, legal differences exist and should be described. To do so, we could add a term that references the category “rest home“ and the French language string “maison de repos”. Two properties with descriptive legal information for respectively Belgium and France could be linked to this French term “maison de repos”. These two properties could then be classified respectively by two categories “Belgium” and “France” with meta category “legal system”.
The CFE uses a dual taxonomical structure to represent the linguistic ontology (see figure 2). The holonym-meronym relation applies to both taxonomies. In the left hand pane the meta category and category taxonomies are displayed. For a selected meta category only the underlying category taxonomy is shown. Information of a selected (meta) category is displayed in the right hand pane. This information includes the list of (multilingual) terminology, the category relations, the category properties and the category classification. In each panel the information may be edited by means of popup menus in combination with additional information windows. For a selected term the term properties, the term relations, and the term classification may for instance be edited in the Term-window.
The (meta) categories in the meta category tree and the category tree may also be displayed in a different language. For this the display language on the toolbar at the top should be selected.
Apart from the selected Categorisation Framework tab there is a Relations tab and a Terminology tab. On the Relations tab bi-directional relations can be managed. On the Terminology tab the terminological database may be queried.

4. Use Case

The CFE was developed and first used during the PoCeHRMOM-project. In part 4.1 we will first describe this project in general. In part 4.2 we will then describe the implementation and role of the CFE during the project.

4.1 The PoCeHRMOM project

This was a joint project of the Centrum voor Vaktaal en Communicatie (CVC) and the Semantics Technology and Applications Research Laboratory (VUB STARLab). The goal of the project was to support e-HRM for SMEs using Semantic Web technology.
During the project an online platform, named Profile Compiler (Tang et al., 2008), was developed on which companies and organisations could create job profiles. A job profile typically contains a list of required competencies and qualifications, together with a list of tasks a job candidate should be able to carry out. The Profile Compiler platform facilitates the creation of job profiles by providing users with a list of occupations to select from. The list of competencies, qualifications and tasks normally associated with that occupation will then be proposed. The user may decide which competencies, qualifications and tasks he wants to include in the job profile. He may also decide to combine the information of multiple occupations or manually add competencies, qualifications and tasks to the customised job profile. The platform also contains a list of competence levels - based on European classification standards - that may be selected to specify the required level of proficiency for a competency. The result of the described actions is a customised job profile the user may save, in the language (EN, FR, NL) of his choice, as a text document for further processing.
More information about the project may be found at the website: http://cvc.ehb.be/PoCeHRMOM/.

4.2 Implementation of the CFE

To materialise the software platform, a multilingual (EN, FR, NL) terminological database was developed describing competencies, competence levels, occupations, qualifications and tasks. The multilingual terminological database was ontologically structured to enable the software platform to manage the terminology on a semantic level. The Profile Compiler platform itself was based on the DOGMA, i.e. Developing Ontology-Grounded Methods and Applications, framework (De Bo et al., 2003) of VUB STARLab.
CVC was responsible for the development of the terminological database whereas VUB STARLab developed the Profile Compiler platform.
Chronologically, we first mutually agreed on the data that was necessary for the Profile Compiler, i.e. competencies, competence levels, occupations, tasks and qualifications. CVC then developed the CF and the CF XML format that could be used to store the ontologically structured terminological information. The CF XML format was used to import the terminological information into the DOGMA framework. Once the CF XML format was tested, the CFE was developed by CVC to manage the CFs. Using the CFE, the required data structure was first implemented at the abstract meta category level. Next CVC started gathering existing information sources containing relevant information. These information

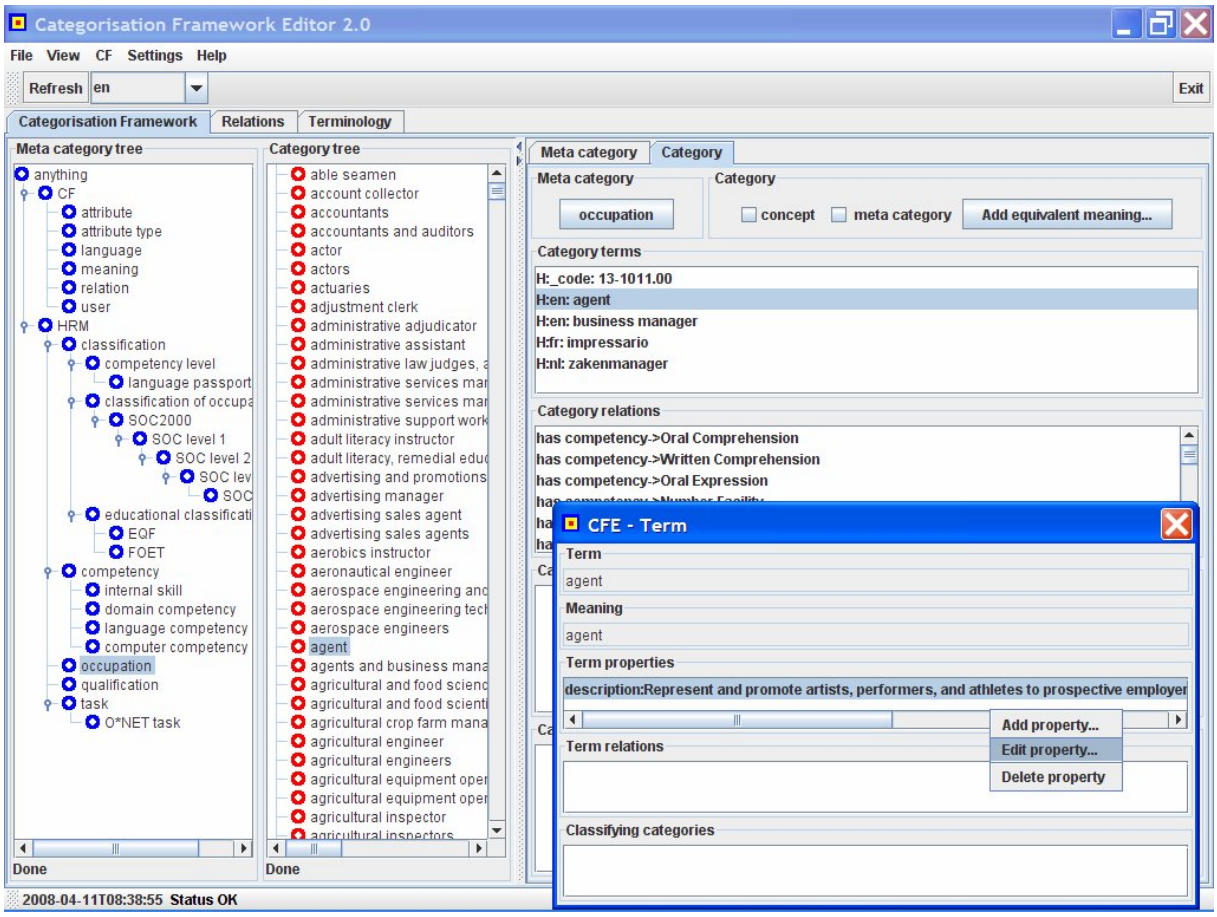


Figure 2: The Categorisation Framework Editor.

sources like the O*NET [9] occupational database, the standard occupational classification (SOC2000), etc. were transformed into CF XML by means of conversion programs written in Java or Perl. The CFE was also used to add translation equivalents and synonyms for the retrieved information where applicable.

In the end the resulting CF XML files were imported in the DOGMA server. The terminological information could then be used on the Profile Compiler platform.

5. Related and Future Work

Although we decided to develop a new data structure and software tool to handle ontologically structured multilingual terminological information, we based ourselves on the works of many previous linguistic, terminological and ontological studies. The insight of de Saussure (1916) that a linguistic sign is a combination of both a signifiant, i.e. symbol, and a signifié, i.e. meaning, is illustrated in figure 1. In the CF a term is a dual structure with reference to both a language string (the symbol) and a meaning item. In semiotics (Chandler, 2006) this idea was further developed to address the phenomenological world by means of language and mental ideas. The semiotic triangle (Ogden & Richards, 1923) distinguishes between referent, thought or reference, and symbol (see figure 1). A referent is any part of the objective world, both that part that really exists and that part that can be imagined. Traditionally a reference is part of our internal reality and only exists in our thoughts. In order to communicate, we need a designator or term that designates the reference and thus denotes the referent. Such a term is the representation of the reference by a symbol. The term is related to the social/normative world as terminology is part of our world of linguistic tradition and norms. The process of communication thus concerns all three types of worlds (Schoop, 2004). In a CF meaning items or references are however elements of a formalised linguistic ontology. There meaning content may thus be shared. On the one hand the CF may help humans to grasp the meaning of terminology, on the other hand applications may use the CF structure to 'understand' terminology. Understanding should here be interpreted in the sense that a software application could respond to user queries containing predefined terminology in an intelligent manner. For example if the user changes the display language of the CFE, the tool may use the terminology in the requested language to represent the (meta) category taxonomy. For this purpose an ontological commitment must be implemented in the software application or agent. We say that an agent commits to an ontology if its observable actions are consistent with the definitions in the ontology (Gruber, 1993).

[9] O*NET OnLine is available at http://online.onetcenter.org/

Like in Frame Semantics (Fillmore 1985; Fillmore and Atkins 1992) terms in the CF are also defined by the semantic relations between there corresponding meaning items, this in addition to the use of descriptive properties, classification and the use of multilingual terminology. In Frame Semantics a separate frame is however constructed for each term whereas the CF could be regarded as a single semantic network.

At the beginning of the PoCeHRMOM-project we tested the terminology management system OntoTerm of Antonio Moreno (Moreno Ortiz, 2000). This tool is comparable to the CFE but since it was not open-source software and had no application programming interface we did not wish to use it for the project.

During the development of the CFE we were inspired by the frame-based Protégé ontology editor (Genari et al., 2002). Meta categories, categories, and meaning relations in the CFE could be viewed respectively as classes, instances, and slots in Protégé. However, in the CFE (meta) categories are but aspects of the underlying meaning items. A meta category could thus also be a category. In the CFE terms also play a more central role. They are used to represent and identify meaning items, may be assigned term properties, term relations and may be classified. The dual taxonomical hierarchy of meta categories and categories in the CFE is also different from the single generic-specific class hierarchy in Protégé. The dual taxonomical hierarchy in the CFE may best be compared to Topic Maps (Pepper, 2000). A meta category could be seen as a topic, a category as an occurrence, and a meaning relation as an association between topics. This may illustrate that the CFE offers a generic interface for knowledge structuring by means of terminology.

In the future we would like to investigate how the CFE could be used during knowledge management projects. More specifically we would like to investigate how a CF may facilitate information retrieval.

6. Conclusion

We believe that the CFE is a practical software tool for the construction of (multilingual) ontologically structured terminological databases. Due to the design of the applied data structure, i.e. the Categorisation Framework, the tool supports the automatic merging of heterogeneous and multilingual terminological databases. Especially in a setting where different domain experts and/or terminographers should collaborate this could be a real advantage.

The information in a CF may be exchanged using the CF XML format or by means of the CF application programming interface. These features could facilitate, the information exchange with other applications, and the conversion to and from other data formats.

7. Acknowledgements

This research and development has been funded by the Flanders IWT-TETRA project PoCeHRMOM (IWT-no. : 50115), and the European Lifelong Learning Programme as part of the Live Performance Technics in the European Union project (grant: LLP-LdV-TOI-07-FI-160 813).

8. References

Chandler, Daniel (2006). Semiotics: The Basics (Paperback). N.e edition: Routledge.

De Bo, J., Spyns, P., Meersman, R. (2003). Creating a "DOGMAtic" multilingual ontology infrastructure to support a semantic portal , On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops, In, R. Meersman, Z. Tari et al.,(eds.), p.253-266,LNCS 2889.

Fillmore, Charles (1985). “Frames and the Semantics of Understanding”. Quaderni di Semantica 6(2):222–254.

Fillmore, Charles and Atkins, Beryl (1992). “Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors”. In: Lehrer, Adrienne and Kittay, Eva Feder (eds.). Frames, Fields and Contrasts. Hillsdale, NJ: Lawrence Erlbaum Associates, 75-102.

Gennari, J., Musen, M. A., Fergerson, R. W., Grosso, W. E. Crubezy, M., Eriksson, H., Noy, N. F., Tu, S. W. (2002). The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. Stanford Biomedical Informatics, Conference proceedings

Gruber, Thomas R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In Formal Ontology in Conceptual Analysis and Knowledge Representation, edited by Nicola Guarino and Roberto Poli, Kluwer Academic Publishers, in press. Substantial revision of paper presented at the International Workshop on Formal Ontology, March, 1993, Padova, Italy. Available as Technical Report KSL 93-04, Knowledge Systems. Laboratory, Stanford University.

Kerremans, Koen (2004). "Categorisation Frameworks in Termontography". In: Temmerman, Rita and Knops, U. (eds.) The Translation of Domain Specific Languages and Multilingual Terminology Management. Linguistica Antverpiensia New Series 3/2004. Antwerp: Hogeschool Antwerpen, p. 263-277.

Ogden, Charles Kay and Richards, Ivor Armstrong (1923). The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism. London: Kegan Paul, Trench, Trubner.

Moreno Ortiz, A. (2000) “Managing Conceptual and Terminological Information in a User-friendly Environment”. Proceedings of OntoLex 2000. Workshop on Ontologies and LexicalKnowledge Bases.

Pepper, Steve (2000). The TAO of Topic Maps, Proceedings of XML Europe 2000, GCA, online at http://www.ontopia.net/topicmaps/materials/tao.pdf.

Saussure, F. de (1916). Cours de linguistique générale, Paris, Payot.

Schoop, Mareike (2004). Proceedings of the 9th International Working Conference on the Language-Action. Perspective on Communication Modelling (LAP 2004). Rutgers University, The State University of New Jersey, New Brunswick, NJ, USA, June 2-3, 2004 (M. Aakhus, M. Lind, eds.).

Temmerman, Rita, Kerremans, Koen (2003).

"Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description". In: Hajicová, E., Kotešovcová, A., Mírovský, J. (eds.), Proceedings of CIL17, Matfyzpress, MFF UK (CD-ROM). Prague, Czech Republic.

Temmerman, R. (2000). Towards New Ways of Terminology Description. The sociocognitive approach. Amsterdam/Philadelphia: John Benjamins.

Vossen, Piek (1998). Introduction to EuroWordNet. Computers and the Humanities: Springer.

Tang, Yan, Christiaens, Stijn, Kerremans, Koen (2008). Profile Compiler: Ontology-Based Multilingual Online Services to Support Collaborative Decision Making. Proceedings of the IEEE International Conference on Research Challenges in Information Science, June 3-6, Marrakech, Morocco.