The Development of a Multilingual Collocation Dictionary

Sylviane Cardey
Centre Tesnière
Université de Franche-Comté
France
sylviane.cardey@univ-fcomte.fr

Rosita Chan
Centre Tesnière
&
University of Panama
chan.rosita@hotmail.com

Peter Greenfield
Centre Tesnière
Université de Franche-Comté
France
peter.greenfield@univ-fcomte.fr

Abstract

In this paper we discuss the development of a multilingual collocation dictionary for translation purposes. By 'collocation' we mean not only set or fixed expressions including idioms, simple co-occurrences of items and metaphorical uses, but also translators' paraphrases. We approach this problem from two directions. Firstly we identify certain linguistic phenomena and lexicographical requirements that need to be respected in the development of such dictionaries. The second and other direction concerns the development of such dictionaries in which linguistic phenomena and lexicographic attributes are themselves a means of access to the collocations. The linguistic phenomena and lexicographical requirements concern variously placing the sense of collocations rather than headwords or other access methods at the centre of interest, together with collocation synonymy and translation equivalence, polysemy and non-reversibility of the lexis, and other more lexicographic properties such as varieties of language and regionalisms, and types of translation.

1 Introduction

In work with developing multilingual collocation based dictionaries for translation purposes across a wide variety of domains (Cardey and Greenfield, 1999; Chan 2005) various interesting linguistic phenomena and lexicographic requirements have been observed. In the context of such dictionaries, by the term collocation we include not only set or fixed expressions (Moon, 1995, Tables 1.1 and 1.2, pp.19-20) including idioms, simple co-occurrences of items (*plane ticket*) and metaphorical uses (*spill the beans*), but also, as we will show, translators' paraphrases where these are needed. Linguistic phenomena include ones concerning sense (for example synonymy and translation equivalence, polysemy and non-reversibility). Lexicographical requirements include for example the requirement (for consistency purposes amongst others) that the collocation (as article) be the centre of interest rather than the headword(s) whose role is one of access to the collocations. This is principally because the object of such dictionaries should be based on inter-lingual collocation sense group correspondence, translation of headwords being essentially incidental. Another way to view this is that if what we wish to model is a dictionary of senses, these senses are expressed by interpretations in the form of collocations. However, difficulties are engendered with this approach. For example, headwords are typically canonical in form whilst their corresponding lexical units in collocations can be variants (for example inflected or be derivations). Furthermore, in reality the definition of a collocation structure for lexicographic purposes can itself be complex, for example to cater for or indicate information such as inflected forms, synonymy and translation equivalence, grammatical labelling and comments (Gaveiro, 1998, pp. 26 - 27, 64 - 65).

More recently, our interest has been concerned with how to develop such multilingual collocation dictionaries including access to collocations based on linguistic phenomena as well as by headwords (a headword can only be a single word, even for idioms) (Chan, 2005) where the issues of particular cases at the semantic level and at the grammatical level are important. Here

Proceedings of the Workshop on Multilingual Language Resources and Interoperability, pages 32–39,
Sydney, July 2006.

the access to collocations can be by posing a problem; one can ask for those collocations which present a problem of article for example.

The linguistic phenomena and lexicographic requirements are ones that are candidates for modelling such dictionaries using formal methods, for example using the Z formal specification language (Spivey, 1992), the impetus being that certain domains in which such dictionaries are used are safety critical in nature. This has resulted in work in respect of the state invariants peculiar to specialised multilingual collocation based dictionaries (Greenfield, 1998a; Greenfield, 1998b).

In response to these various observed linguistic phenomena and lexicographical requirements, the MultiCoDiCT (Multilingual Collocation Dictionary System Centre Tesnière) system was developed as a research aid tool for multilingual lexicographic research (Greenfield et al., 1999; Greenfield, 2003). The basic model underpinning MultiCoDiCT dictionaries reposes on the concept of the collocation sense group as a means to ensure integrity and consistent access to the collocations. In this model a collocation in a language appears only once, whereas in conventional dictionary models it is the headword in a language that appears only once. This constraint leads us to generality; not only do we obtain reversibility of translation with no extra effort, we obtain non-reversibility of the lexis where this happens to be the case. Furthermore, headword access to a collocation also provides direct access to the other collocations in the dictionary with an equivalent sense (or senses for polysemic collocations).

More recently, work on linguistic phenomena and lexicographic attributes based access to collocations (as well as headword access) has resulted in a prototype system using an algorithmic approach (Chan, 2005) using the Studygram system (Cardey and Greenfield, 1992).

In the paper we first review the linguistic phenomena and lexicographic requirements that we have discerned for such multilingual collocation dictionaries. We then discuss the development of such dictionaries in which the linguistic phenomena and lexicographic attributes are themselves a means of access to the collocations. Finally, in the conclusion we show how Studygram and MultiCoDiCT can be integrated in order to provide a more general approach for the access to such multilingual collocation dictionaries.

2 Linguistic phenomena and lexicographic requirements

In the context of lexicographical research, collocations as articles in multilingual dictionaries present various linguistic phenomena and lexicographic requirements which are sufficiently generic but also sufficiently important lexicographically as to warrant some generalised support. The various phenomena and requirements are illustrated in this section by the essentially traditional headword access method to collocations as provided by the MultiCoDiCT system.

The linguistic phenomena concern synonymy, polysemy and non-reversibility of the lexis in translation. For example synonymy is indicated by more than one collocation having the same sense equivalence variously in the source language or in the target language (in the illustrations that follow the source language is on the left and the target language is on the right); see Figures 1 and 2.

Spanish	French
Headword	
boleto	billet
Collocations	
billete de avión (Spain) boleto de avión (Americanism)	billet d'avion

Figure 1. Synonymy in the source language

French	Spanish
Headword	
billet	billete(Spain) boleto(Americanism)
Collocations	
billet d'avion	billete de avión (Spain) boleto de avión (Americanism)

Figure 2. Synonymy in the target language

In the above two examples, the Spanish collocations include annotations indicating regionalisms such as *(Spain)* (Chan, 1999). We say that collocations in the same or different languages which are equivalent in that they have the same sense are members of the same *sense group*. In the above examples we can also observe various lexicographical requirements such as headwords and the use of structured annotations to display the regionalism information.

Polysemy is indicated by the presence of translations with different senses, that is, where a collocation is the member of more than one sense group. The example that we use is drawn from an archtypical bilingual dictionary (Werthe-Hengsanga, 2001) of Thai-French image expres-

sions in the zoological domain with the particularity that the types of translation are shown by lexicographic annotations as follows:

- Eq equivalent – supplied, provided that an equivalent can be found
- LT literal translation – word for word
- AS analytical sense – literal translation reformulated in correct French
- FS functional sense – the true sense of the translated collocation

The Thai is shown here by means of a phonetic transcription using ASCII characters (which in fact does not provide an adequate cover but this matter is not pursued here). An example of a polysemic Thai collocation with 3 functional senses (FS) is shown in Figure 3.

Thai	**French**
Headword	
hmu:	cochon, porc
Collocations	
j?:n hmu: me: w: (TP)	**sense 1:** donnant donnant (Eq) tendre porc<n(m,s)> tendre chat<n(m,s)> (LT) l'un tend son cochon<n(m,s)> l'autre son chat<n(m,s)> (AS) contre une chose, une prestation équivalente à ce qu'on donne soi-même (FS) **sense 2:** donnant donnant (Eq) tendre porc<n(m,s)> tendre chat<n(m,s)> (LT) l'un tend son cochon<n(m,s)> l'autre son chat<n(m,s)> (AS) prendre son dû séance tenante dans une transaction (FS) **sense 3:** donnant donnant (Eq) tendre porc<n(m,s)> tendre chat<n(m,s)> (LT) l'un tend son cochon<n(m,s)> l'autre son chat<n(m,s)> (AS) vendre et acheter comptant (FS)

Figure 3. Polysemy illustrated by a Thai collocation with 3 functional senses (FS)

The linguistic phenomenon 'non reversibility of the lexis' is illustrated by the example shown in Figure 4.

French	**English**
Headword	
antécédents	'medical history'
Collocations	
antécédents du patient	patient history
antécédents médicaux	medical history

English	**French**
Headword	
history	–
Collocations	
patient history	antécédents du patient
medical history	antécédents médicaux

Figure 4. Illustration of non reversibility of the lexis

In this dictionary which is restricted to the domain of clinical research (Gavieiro 1998), even though there is a translation of the French headword *antécédents* by an English collocation '*medical history*' (printed between quotes to indicate it to be a collocation rather than a headword in the target language), this is not the case for the inverse sense for the English headword *history*. Being a dictionary of collocations, the translation of *history* as a headword has no place in such a domain specific dictionary. On the contrary, English collocations containing the headword *history* have their place, they are translated to French.

Lexicographic requirements can be divided into those which concern the functionality offered by the dictionary (for example, as we have already seen, the use of annotations for various purposes) and those which concern the organisation and integrity of the dictionary.

The functionality offered by such a dictionary includes the method of access to collocations as articles, the presentation of the articles in order to display any of the linguistic phenomena present (as has been illustrated by the examples above concerning synonymy, polysemy and non-reversibility of the lexis in translation), and the organisation and provision of lexicographical annotations.

For the access to collocations as articles this can be as in conventional paper dictionaries by means of headwords, typically in alphabetic order. A headword is an individual lexical unit whose primary purpose is to provide a means of access to a collocation. In the MultiCoDiCT system a headword is never the whole collocation even for a fully fixed expression. A given headword can access several collocations (as is illustrated in Figure 4) and in like manner,

a collocation can be accessed by many headwords. This can be seen for the collocation *antécédents médicaux* with headwords *antécédents* (Figure 4) and *médical* (Figure 5).

French	English
Headword	
médical	medical
Collocations	
antécédents médicaux	medical history

Figure 5. Variant form of headword in the context of the collocation

The headwords of a collocation require to be specified; in the MultiCoDiCT system this is done explicitly by the lexicographer. Because of inflexional and derivational morphology, the headwords are typically in a canonical form, whilst the forms in the collocations can be variants; Figure 5 illustrates this for the French headword '*médical*' which takes the variant form '*médicaux*' in the French collocation. In Figure 4, the case of the headword *antécédents* (nominally a 'variant' (plural) of the canonical form *antécédent*) is atypical, the lexicographical choice of the form of the headword here being due to *antécédents* being a short form of *antécédents médicaux*. Thus in the organisation of the dictionary there must be, as is the case in the MultiCoDiCT system, a mapping between headwords in their canonical form and their 'presence' in collocations.

With annotations such as grammatical function (already shown in Figure 3) even the linguistic phenomenon of grammatical variation can be accounted for, as shown in Figure 6.

French	Spanish
Headword	
aérogare	terminal<n(m,s)> 'estación<n(f,s)> terminal <adj(f,s)>'

Figure 6. Illustration of grammatical variation

In the case of synonyms or polysemic equivalences, a given word can 'change' its grammatical role. In the first synonymic equivalence in the example in Figure 6, the Spanish word *terminal* is a noun whilst in the second it has as grammatical function adjective because the word *estación* has as role a noun. It should be noted that for the two grammatical functions of the word form *terminal*, in the Spanish lexis there is only one headword for *terminal*.

We now turn to the phenomena which have an impact on the organisation and integrity of such a dictionary and thus its underlying model and how this has been achieved in the MultiCoDiCT system. We must deal with variously collocations, headwords and annotations and the various interrelations between these such as sense groups and furthermore the relation between headwords and collocations, all these in a multilingual context. There must necessarily be some means to ensure the integrity of these various data items and the relations between them.

The model that underpins the MultiCoDiCT system is based on:

- firstly the sense group to which one or more collocations across one or more languages is associated in being sense equivalent (a sense group is no more than such an association),
- secondly the languages, to each of which collocations are uniquely associated,
- thirdly the collocations and
- fourthly the headwords, which in the MultiCoDiCT system are the only way to access directly a collocation and its sense equivalences (synonyms and translations). (Acces to collocations by means of linguistic phenomena is discussed in the next section.)

In respect of annotations, the underlying model allows these to be added at the level of sense group, collocation (for example regionalism) or collocation lexical item (for example grammatical category)

As far as the collocations which are the members of a sense group are concerned, these can be viewed orthogonally over two dimensions. One dimension involves the language and here too one or more languages may have a special status, such as Latin in dictionaries of flora and fauna which we address in the next section of the paper. The other dimension concerns the nature of collocations. Here we can type collocations as either being 'true' collocations in terms of the linguist's view, or, collocations which are translators' paraphrases such as for example translation types as we have already discussed and shown (Figure 3).

3 Linguistic phenomena and lexicographic attributes as a means of access

In this section we consider linguistic phenomena and by extension lexical attributes such as annotations of linguistic phenomena as themselves a means of access to the collocations in multilingual collocation dictionaries. We illustrate this approach by describing a bilingual dictionary of tourism that has been developed with this means of access in mind.

This dictionary involves the differences between French-Spanish-French translations found in ecological, fluvial and cultural (historical and religious) tourism corpora (Mendez, 1993; Rigole and Langlois, 2002). When translating the corpora, we noticed the presence of varieties of languages, such as Latin American Spanish and common Spanish (that is the Spanish of Spain) and of regionalisms; for example, in the case of Panama whose animal and plant specimen names were used only in that country and not in other Latin American countries, see Figure 7 (Chan 2005).

Common Spanish names	corozo	agutí
PANAMA	corozo	ñeque
BOLIVIA	totai	-
CUBA	-	jutía mocha
MEXICO	-	cotuza
VENEZUELA	corozo	zuliano de grupa negra
French translation	acrocome / coyol / noix de coyol	agouti

Figure 7. The presence of varieties of languages

We also found cases of linguistic phenomena at the semantic level, such as Americanisms, Anglicisms, non-reversible equivalents, etc. To handle these various observations we developed an algorithmic dictionary access method in order to provide access to the relevant collocations and translations. Our overall algorithm (see Figure 8) is itself composed of three principle sub-dictionaries:

a. French-Spanish-French equivalences (537 words),
b. particular cases at the semantic level (1146 words) and
c. particular cases at the grammatical level (291 sub-conditions).

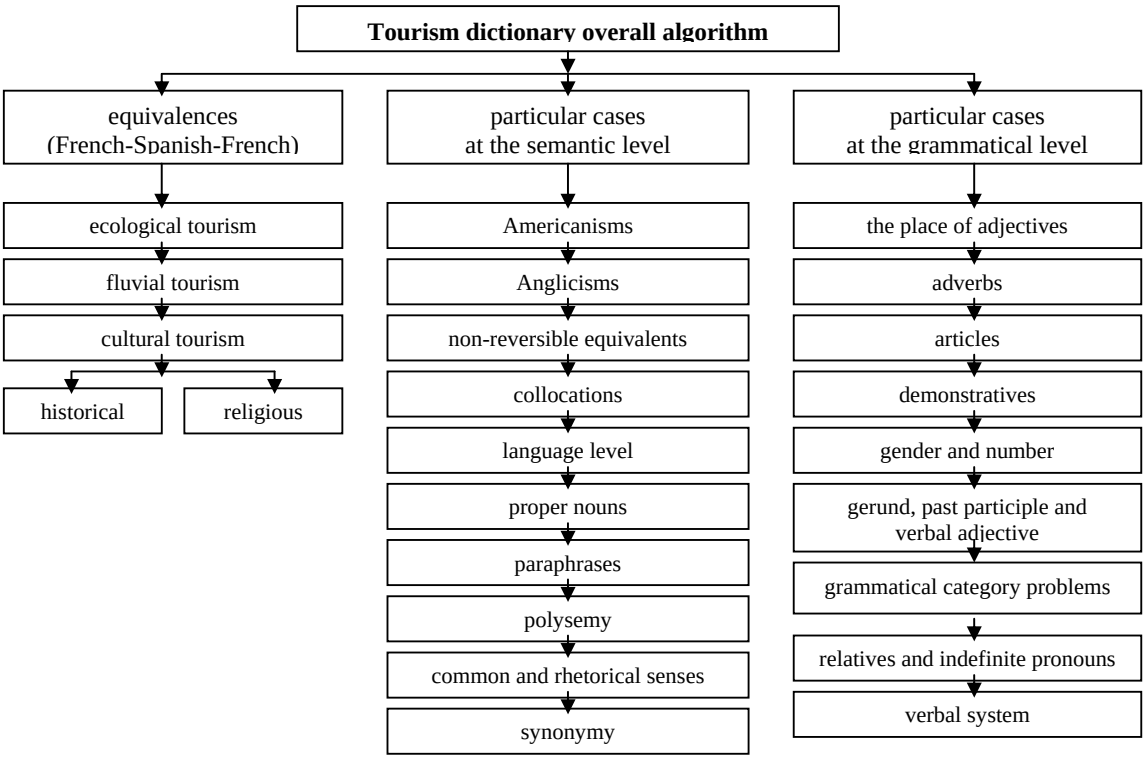

Figure 8. Dictionary access algorithm

The algorithm has a maximum of eight levels where the existence of other sub-dictionaries (or sub-algorithms) is possible inside of each dictionary, which itself can be consulted independently or dependently. In other words, the overall algorithm includes several mini-algorithms and mini-dictionaries.

At the start of each consultation, the words belonging to a given dictionary are presented in the form of a list arranged in alphabetical order so that the user can save time.

We now discuss these three specific sub-dictionaries.

The first sub-dictionary concerns equivalences which are provided in the French-Spanish-French languages and which are classified according to topic. The sub-field cultural tourism presents for example historical and religious tourism as sub-sub-fields.

The second sub-dictionary concerns particular cases at the semantic level, the terms of the dictionary of the Panamanian fauna, for example, are joined together by class such as: insects, mammals, birds and reptiles. The user can check starting from:

- French to obtain the equivalences in Spanish of Panama and common Spanish;
- French to obtain the equivalences in common Spanish and Latin;
- Panamanian Spanish to obtain the equivalences in common Spanish;
- common Spanish to obtain the equivalences in Panamanian Spanish;
- Panamanian Spanish and common Spanish to obtain the equivalences in French and Latin;
- Latin to obtain the equivalences in French, common Spanish and Panamanian Spanish.

At the outset we had the intention to develop a bilingual dictionary. However, we included Latin in the dictionary, since, when translating the Spanish corpora to French, we noticed that the names of the flora and fauna belonged to a specialised lexicon and that most of these names constituted regional variations. Thus, we had to look for the scientific name (coming from Latin), then the common Spanish name in bibliographical documents, monolingual dictionaries or on Internet sites dedicated to these fields and finally, to look for the French translation in general bilingual dictionaries (Spanish-French) and on zoological and botanical websites in order to validate the equivalences.

We did not consider the variants of other Latin-American countries because in order to do so it would have been necessary to undertake an intensive research exercise into the matter and to have had the terms checked by specialists in the field studied.

The third and last sub-dictionary deals with grammatical findings. It is not only composed of words but grammatical rules and also examples in order to illustrate the different cases. For this reason, we do not mention the quantity of words in the dictionary but rather the number of sub-conditions in the algorithm.

The algorithm that we have developed is interactively interpretable by the Studygram system (Cardey and Greenfield, 1992) which also provides the user interface. To illustrate the trace of a user access using our prototype system with the dictionary access algorithm illustrated in Figure 8, we take as entry the French collocation '*amazone à front rouge*' and where we are interested in the equivalences sub-dictionary and the particular cases sub-dictionary (see Figure 9).

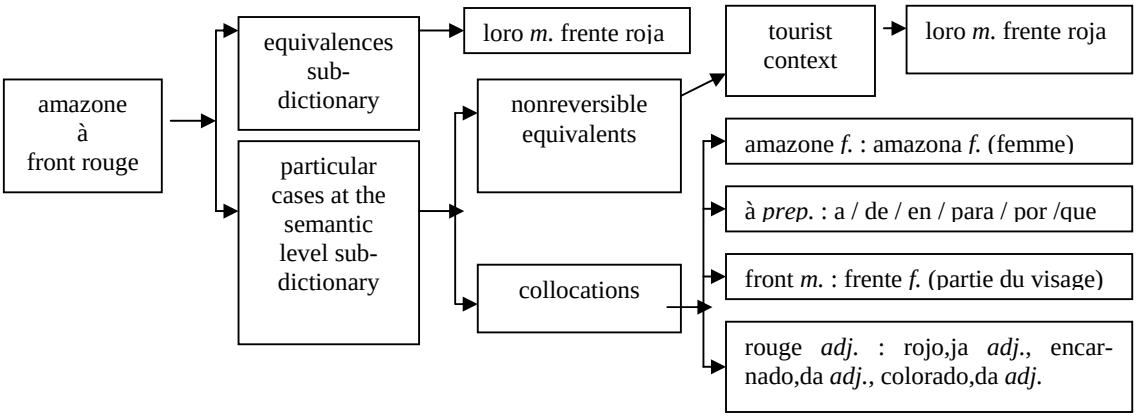

Figure 9. Trace of a user access with as entry the French collocation '*amazone à front rouge*'

4 Conclusion

We have presented the essential linguistic phenomena and the lexicographic requirements that we have discerned to be useful in the development of multilingual collocation dictionaries. By 'collocation' we mean not only set or fixed expressions including idioms, simple co-occurrences of items and metaphorical uses, but also translators' paraphrases. The basic model for such dictionaries as exemplified in the MultiCoDiCT system reposes on the concept of the collocation sense group as a means to ensure integrity and access. We have presented a novel access method to such dictionaries using the Studygram system which in essence provides access also based on much of the very linguistic phenomena and related lexicographical attributes that we have previously discerned and thus enabling access to collocations by posing a problem.

We conclude in showing how Studygram and MultiCoDiCT can be integrated in order to provide a more general approach for the access to such multilingual collocation dictionaries. In this approach, Studygram would provide the user interface and problem solving capability as described in section 3 and MultiCoDiCT would act as a lexical resources server.

At one level this approach would involve the essentially technical matter of standardising the mutual call mechanism (operational semantics) between Studygram and MultiCoDiCT. The Studygram system in any case supports algorithm solutions (called operations) which can be procedure calls, which in this context would be to MultiCoDiCT.

At another level this approach would involve formalising and standardising the linguistic and lexicographic terminology shared by the two systems. This level is thus concerned with including the lexicographical needs in the computational model. In respect of the semantics of the MultiCoDiCT component, the model underpinning MultiCoDiCT could be extended in a simple fashion to support explicitly the provision of linguistic 'headwords' involving the intrinsically modelled linguistic phenomena of synonymy and translation equivalences, polysemy and non-reversibility of the lexis. Access by conventional headwords is in any case already supported. The same MultiCoDiCT model provides annotation structures attached to the sense group, to the collocation and to the collocation lexical item. However the semantics of the annotation content is the lexicographer's and thus would involve an agreed semantics between the MuliCoDiCT and Stydygram components including the algorithm content concerning the machine interpretation of such annotation contents and lexicographic attributes.

References

Cardey, S., Greenfield P. 1992. The `Studygram' Natural Language Morphology System: A First Step to an Intelligent Tutoring Platform for Natural Language Morphology in *Proceedings of the UMIST ICALL workshop*, 42-59, published by CTI Centre for Modern Languages, University of Hull, UK

Cardey, S., Greenfield, P. 1999. Computerised set expression dictionaries : design and analysis, *Symposium on Contrastive Linguistics and Translation Studies* (Université Catholique de Louvain, Louvain-la-Neuve, Belgique) 5-6 February 1999. In Lexis in Contrast (B. Altenberg & S. Granger eds.), Benjamins.

Chan, R., 2005, El diccionario de la flora y fauna panameña: propuesta de algoritmo para la solución de problemas de traducción de español-francés, *IX Simposio Internacional de Comunicación Social, Santiago de Cuba*, 24-28 de enero 2005, Actas I, 389-393.

Gaveiro, E., 1998, Elaboration d'un Prototype de Dictionnaire Informatisé Français-Anglais / Anglais-Français. Application au Domaine de la Recherche Clinique, Mémoire de D.E.A., Université de Franche-Comté, France.

Greenfield, P. 1998a. L'espace de l'état et les invariants de l'état des dictionnaires terminologiques spécialisés de collocations multilingues, *Actes de la 1ère Rencontre Linguistique Méditerranéenne, Le Figement Lexical*, Tunis, les 17-18 et 19 September 1998, 271-283.

Greenfield, P. 1998b. Invariants in multilingual terminological dictionaries, *BULAG N° 3, ISBN 2-913322-11-5, Presses Universitaires Franc-Comtoises*, 1998, 111-121.

Greenfield, P. 2003. Le rôle de l'informatique dans le traitement et l'enseignement des langues, *Actes du Congrès international : Journées linguistiques franco-asiatiques*, Naresuan University, Phitsanulok, Thaïlande, 20-22 August 2003, 69-84.

Greenfield, P., Cardey, S., Achèche, S., Chan Ng, R., Galliot, J., Gavieiro, E., Morgadinho, H., Petit, E. 1999. Conception de systèmes de dictionnaires de collocations multilingues, le projet MultiCoDiCT, *Actes du Colloque international VIème Journées scientifiques du Réseau thématique de l'AUF Lexi-*

cologie, Terminologie, Traduction, Beyrouth, 11-13 November 1999, 103-113.

Mendez, E. 1993. *Los roedores de Panamá*. Panamá, Laboratorio Conmemorativo Gorgas, pp.59-64 and pp.281-286.

Moon, R. 1998. *Fixed expressions and idioms in English, a corpus-based approach*. Clarendon Press, Oxford. ISBN 0-19823614-X.

Rigole, M., Langlois, C-V. 2002. *Panamá. Guides de voyage ULYSSE*. 4th edition. Canada, Guides de voyage Ulysse inc., 333p.

Spivey, J.M. 1992. *The Z Notation*. Prentice Hall, ISBN 0-13-978529-9.

Werthe-Hengsanga, V. 2001. *Etude de la traduction automatique en français des expressions imagées de la langue thaï (domaine animalier)*. DEA, Sciences du langage, Université de Franche-Comté, Besançon, France.