Interlinguistic Divergences in Papillon Multilingual Dictionary

Mathieu MANGEOT
National Institute of Informatics
2-1-2-1913, Hitotsubashi
Chiyoda-ku Tokyo 101-8430, Japan

mangeot@nii.ac.jp

Kyoko KURODA
Shimane Prefectural Women's College
7-24-2 Hamanogi
Matsue Shimane 690-0015 Japan
& Centre de recherche en linguistique Lucien Tesniere 30, rue Megevand, 25030
Besançon Cedex, France

kyoko.kuroda@nifty.ne.jp

Abstract

The Papillon project aims at building a multilingual pivot dictionary by voluntary contributors in collaboration on the Internet. The project started as a French-Japanese cooperation. The languages on development are English, French, German, Japanese, Lao, Malay, Thai and Vietnamese. We will discuss about how to solve the problems raised by the divergence of the predicates' argument structures in a multilingual context. If the entry (word sense) is a predicate (e.g.: "murder"), the semantic formula describing the entry will represent this predicate (e.g.: "murder of the individual X by the individual Y"). However, the number and the position of the arguments may vary a lot among languages. For example, there is a position shift in the arguments between the English predicate "X1 misses Y1" and its French equivalent "X2 manque à Y2" with X1=Y2 and Y1=X2. We propose a way to note this divergence in order to translate structures correctly.

1 Introduction

The Papillon project aims at creating a multilingual lexical database with, among others, English, French, Japanese, Malay, Lao, Thai and Vietnamese. The access is free as long as there is no commercial use (open source license). Our project is open to the collaboration of any people interested in the human languages. The macrostructure of the dictionary is composed of a monolingual volume for each language and a pivot volume containing the interlingual links linking word senses in the monolingual volumes [17], [18]. The microstructure of the entries is based on the combinatorial lexicography of the meaning-text theory [10], [11], [12]. The lexical database is built on a platform for manipulating dictionaries [9] available mainly on the Web. Everybody can consult the existing data and then correct, complete this data.

Until now, the Papillon dictionary proposed only interlingual links of the same type: lexies, examples or idioms. It was not possible to represent the divergence of the argument structure of lexies/predicates of different languages with the Papillon formalism. This issue was mentioned by several people during informal discussions on Papillon project.

This problem of the interlinguistic divergence is well known and has been already broadly covered in the litterature [14]. The specificities of this paper are to address this problem in a relatively new framework. First of all, the macrostructure of the dictionary is multilingual, and there is no limit about the number of the possible covered languages. Thus all the solutions concerning bilingual dictionaries cannot be applied here. We must find a solution that is independent from the languages and that does not explode when a new language is added.

Second point, we have to cope with the specific microstructure of Papillon which is based on combinatorial lexicography, part of the meaning-text theory. The solution must be in accordance with this theory.In this paper, we propose a modification of the formalism in order to represent these divergences. We also hope to open the discussion on this issue and invite the readers of this paper to send us remarks.

2 Microstructure of the Papillon Dictionary

2.1 Monolingual Entries Microstructure

The basic lexical unit of a monolingual volume entry of Papillon Dictionary is the lexie (a set of signs with the same lexical meaning) A lexie is either a lexeme, or a full idiom.

The usual convention in order to distinguish between a lexie, a character string or a word-form is to write the name of the lexie in capital letters. For example, we write the lexie MEURTRE (murder) to talk about the lexie with the headword “meurtre”. This convention seems inadequate when using a multilingual dictionary because lots of languages like Japanese, Chinese or Thai do no have any distinction between capital and small letters. We adopt here a new convention consisting in using the unique identifier of the lexie (in the Papillon database) to speak about it. This identifier is built by adding the ISO 639-2/T 3 letter code of the language of the lexie (e.g.: fra for French, eng for English, jpn for Japanese, etc.), followed by a dot ‘.’, followed by the lexie headword, optionally followed by an underscore ‘_’ and an integer in roman numbers representing a unique identifier if the lexie is part of an homographic vocable and then followed by an integer reprensenting the unique identifier of the lexie in the vocable.

We use the dot (character ’.’) as a separator because it is accepted in XML identifiers unlike ‘$’ and ‘#’. The headword of the lexie must be written in latin characters or in ideograms. For example, the Japanese lexie SATSUJIN will have the following identifier lexie `jpn.殺人.1` (satsujin) (murder). For languages using non-latin alphabets like Thai or Arabic, we must decide of a transcription in latin characters. For more precisions, see the production rules of a Name in the XML 1.0 recommendation by W3C. With the new convention, we will write the lexie MEURTRE fra.meurtre.1. In the case of homographic vocables, for example the French “campagne (électorale)” (campaign) and “campagne (où broutent les vaches)” (countryside), we will write: `fra.campagne_I.1` and `fra.campagne_II.1`. We are not specialist about conventions, this is why we welcome any proposal favorably.

The structure of the Papillon database monolingual volumes is directly taken from the DiCo project database. The semantic formula of a lexie (semantic characterization of the lexie and its arguments) is made of a semantic label and the predicate defining the lexie. In the case of verbs, the structure of the arguments will be represented by this predicate. For example, the semantic formula corresponding to the French verb “rêver” (to dream) is the following: « personne X rêver de Y ». X and Y are the logical arguments of the predicate and also, syntactically, the actors of the verb “rêver”.

The government pattern of the lexie represents the syntactical realization of the predicate arguments. In the case of the last formula, the pattern is the following: « X = I = num; Y = II = num ». X is the first argument of the predicate and is a nominam expression. Y is the second argument and is also a nominal expression.

The main syntagmatic and paradigmatic links between the lexie and other lexies of the same language are encoded with lexical functions. Special lexical functions link the head lexie with typical lexies used to speak about its arguments. For example, a possible noun, even if it is seldom used, of the first argument of `fra.rêver.1` is `fra.dormeur.1` (sleeper), encoded with the lexical function S1: S1(`fra.rêver.1`) = `fra.dormeur.1`.

2.2 Structure of Interlingual Links

At first, the interlingual links were representing only translation links between lexies of different languages. These links are called axies (interlingual acceptions). For example, the French lexie `fra.meurtre.1`, the English lexie `eng.murder.1` and the Japanese lexie `jpn.satsujin.1` are linked by the same axie.

Afterwards, we defined interlingual links for examples (exies) and full idioms (ixies). These interlingual links link examples and full idioms that are contained in the lexies. They do not link other types of lexies. This definition has not been implemented.

3 Divergent Argument Structures Issue

When two verbs of different languages are linked by a translation, the argument structures of the predicates realized by these verbs can match or diverge. When the structures match, the equivalent is easy to find for the one who wants to translate a sentence including one of these verbs in another language. For example, the argument structures of the lexies `fra.manger.1` and `jpn.taberu.1` (to eat) match. They are parallel: « individu X1 manger Y1 » and « 人X2がY2を食べる »(hito X2 ga Y2 wo taberu). The French argument X1 matches the Japanese argument X2, and the same for the Y argument. It is then easy to match when translating a sentence from one language to another. For example, a French beginner in Japanese can translate the French sentence “ Je mange du pain. ” (I eat bread) X = je (I) = 私 (watashi); Y = pain (bread) = パン (pan). In Japanese: 私がパンを食べる (watashi ga pan wo taberu).

When the argument structures diverge (see [3] for a classification of the divergences between French and Japanese), in some cases, the match is impossible to find if one does not know both languages.

3.1 Argument Position Shift

Let’s examine first a simple example of argument position shift between French and English:

Lexie `fra.manquer.1`(X,Y) « individu X manque à individu Y ».
Lexie `eng.miss.1`(X,Y) « individual Y miss individual X».
The arguments X and Y are shifted between French and English.

3.2 Coalescence Divergence

Let's examine next two examples of divergence called coalescence between French and Japanese:

The French lexie fra.rêver.1 is translated into Japanese by the phrase "yume wo miru" that means literally: to see a dream (voir un rêve) (yume = dream, rêve ; miru = to see, voir). The lexie fra.rêver.1 will then not have an exact equivalent in Japanese. Nevertheless, it is perfectly possible to translate a French sentence with the verb "rêver" in Japanese without any meaning loss.

Likely, the French verb "peser" is translated into Japanese by the phrase "taijuu/omosa wo hakaru" that means litteraly: measure the weight (mesurer le poids). The lexie fra.peser.1 (to weight) will not have an exact equivalent in Japanese.

If the Japanese semi-idiom "yume wo miru" and the French lexical unit "rêver" match despite their configuration difference, it is because they have both the same meaning. In other words, this same meaning is verbalized by only one word in one language and by a sequence of words in another. It is then obvious to link these meaning units by an axie. But until now, we cannot represent directly a translation link between a word and a full idiom in the structure of Papillon dictionary.

4 Planned Solutions

4.1 Encoding the Divergence into the Axies

The first idea is to encode that divergence into the axies. It seems rapidly unfeasible because the way to diverge varies among the language pairs. Thus, we would have to represent in the axies the divergences for each language pair: those between Japanese and French, those between Japanese and Malay, those between English and French, etc.

Moreover, the divergences are not so automatic, i.e. operable beyond the languages. It then seems impossible to design a way to represent them regardless of the languages.

At last, axies were first designed as a simple translation link between several lexical units. It is then better to stay in this simple perspective.

4.2 Linking the Lexies with Lexical Functions

For the moment, only the lexies (or more exactly the headwords of these lexies) are linked together by axies. We can only find word-to-word matches between the languages.

If we describe idioms in the lexie and if these idioms have their semantic equivalent in other languages, it would be desirable to link information units other than the headword in order to link idioms with their equivalents, being another idiom or a single lexeme.

Lets take as examples the above pairs: (A) "rêver" ~ "yume wo miru" and (B) "peser" ~ "taijuu/omosa wo hakaru".

(A) "rêver" ~ "yume wo miru"

First, the government pattern of the lexie `jpn.yume.1` is described as follows [12]:

(1) `jpn.yume.1`
X=I N no
Y=II N no
CI: Marie no yume (Marie is the sleeper)
CII: ryokoo no yume (ryokoo (trip) is the image of the dream)

The letters X and Y are the arguments of the predicate corresponding to the lexie.
The latin numbers I, II, etc. denots the deep-syntactic actants.
CI, CII, etc, represent the columns of the government pattern described in the Explanatory and Combinatorial Dictionary.

This lexie is linked to the French lexie `fra.rêve.1`, and the English lexie `eng.dream.1`, etc.
It has also as a lexical function:

Oper1(`jpn.yume.1`) = `jpn.miru.1`

This LF shows that the name "yume" can combine with the verb "miru".

At the same time, the verb "rêver" corresponding to the syntagmatic expression "yume wo miru" is described in the French volume as follows:

(2) `fra.rêver.1`

X=I N
Y=II de N
CII : "rêver d'une personne, d'une chose" means seeing someone or something during the dream.
S0 (fra.rêver.1) = fra.rêve.1
S1 (fra.rêver.1) = fra.dormeur.1

Note: the verb "rêver" corresponds to at least two different lexies. The lexie fra.rêver.1 with the second argument realized by « de N » is translated in Japanese by the syntagmatic expression "yume wo miru". The lexie fra.rêver.2 with the second argument realized by « à N » means "to hope".
If the lexical function Oper1(jpn.yume.1) = jpn.miru.1 could be linked to the French lexie fra.rêver.1, we would be able to establish a direct link between the syntagmatic expression "yume wo miru" and the verb "rêver"

As for the distribution of the actors, we know, thanks to the Oper1 LF numbering, that the grammatical subject of "yume wo miru" is the sleeper. Likely, by the description of the government pattern of the lexie jpn.yume.1, we see that the name can have modifiers « 1=X=N no » and « 2=Y=N no ». Here, the sleeper is already in the position of grammatical subject. Therefore, we can deduce that the modifier of "yume", i.e. the N of « N no yume » is the image that the sleeper is seeing while dreaming.

With the descriptions of the lexie jpn.yume.1 and the lexical function Oper1 (jpn.yume.1) = jpn.miru.1 on the one hand and the lexie fra.rêver.1 on the other hand, we can obtain the correspondence of the lexical usage in the two languages: « X ga Y no yume wo miru » = « X rêve de Y ».

(B) "peser" ~ "taijuu/omosa wo hakaru"

For the pair "Marie pèse l'enfant" and "Marie ga kodomo no taijuu/omosa wo hakaru" (Marie weights the child), let's read first the description of the Japanese lexie jpn.taijuu.1 (weight).

(3) jpn.taijuu.1
X=I N no
CI : N is a name for a 'human' or an 'animal'.

Among the LF, we will find:
QSyn (jpn.taijuu.1) = jpn.omosa.1
Func1 (jpn.taijuu.1) = jpn.aru.1 (this idiom can be linked via an axie to the lexical units "pesant" and "lourd" (heavy))
Real1(jpn.taijuu.1) = jpn.hakaru.1

For the grammatical subject of the verb "hakaru", we will know by the description of the government pattern that a noun (N) designating the individual can come in the position of a subject:

(4) jpn.hakaru.1
X=I N ga
Y=II N wo
Z=III N de
CI: êl (hito = individual)
CII: ìxó çt (doryoukou = parameter) NAGASA (length), OMOSA (weight)... are mandatory except when we can deduce them from the context.
CIII: measure instrument MONOSASHI (ruler), HAKARI (balance) ... ...

We will then have the semantic formula «"hito" X ga "measure instrument" Z de "doryoukou" Y wo hakaru».

The verb "peser" corresponds to two different lexies. The lexie fra.peser.1 like in the sentence "Ce sac pèse 30 kg." (This bag weights 30 kg.) and the lexie fra.peser.2 like in the sentence "Pierre pèse son sac.".
About the lexie fra.peser.2, the pattern is as follows:

(5) fra.peser.2
X=I N
Y=II N
Z=III 1.avec N, 2.à N, 3.dans N
CI: N désigne un individu (individual)
CII: N désigne un objet concret (real object)
CIII-1: balance
CIII-2: trébuchet, bascule (weighing machine)
CIII-3: main (hand)
Sinstr(fra.peser.2) = balance, trébuchet, bascule
Example : Pierre pèse son sac. (Peter weighs his bag)
Thus, we obtain the semantic formula « Individu X peser Y avec Z ».

If it is possible to link the lexical function Real1(jpn.taijuu.1) = jpn.hakaru.1 with the lexie fra.peser.2, we can link the idiomatic expression « (N no) taijuu wo hakaru » and the lexie fra.peser.2. Note that this "hakaru" is put into relation with the verb "mesurer" (to measure) by another axie.

The axies and the information included in the LFs allow one to deduce the correspondence between two entities, “peser” and “taijuu/omosa wo hakaru” and, then, the correspondence of their arguments, « X peser Y » and « X ga Y no taijuu/omosa wo hakaru ».

4.3 Replacing the Axies by Interlingual Lexical Functions

Another solution would consist in replacing the axies by interlingual lexical functions. The link between the lexies fra.rêver.1 and jpn.yume.1 would then be denoted with interlingual lexical functions: S0[jpn](`fra.rêver.1`) = `jpn.yume.1`, QSyn[jpn](`fra.rêve.1`) = `jpn.yume.1`. This solution has the advantage of simplifying the structure of the lexies. It can be applied realistically for the construction of a bilingual dictionary but in the case of a multilingual database, it becomes difficult to manage directs links from a language to another. It is then better in that case to stay with a macrostructure of monolingual volumes interconnected.

5 Chosen Solution

In the last version of the monolingual microstructures XML schema [8], we suppressed the reference from the lexies to the axies in order to first simplify the maintenance and the checking of the interlingual links and also to represent only information pertaining to the language itself. Links between lexies are therefore only managed with axies. A unique identifier is assigned to each lexie when it is imported in the Papillon dictionary. Axies linking to this lexie use this identifier. If the lexie is destroyed, the identifier is not reallocated.

We propose to simplify and generalize the interlingual links system. Instead of having for each linked information unit a different interlingual link (axie, exie, ixie, etc.), we propose to:

1. link lexie headwords instead of entire lexies;
2. declare linkable the following information units: the headword, the lexical functions, the examples and the full idioms;
3. link all the different information units with the same link type: an axie;
4. link heterogeneous information units with only one axie. For example, in the previous case, we would link the lexical function (jpn.yume.1)=jpn.miru.1 with the lexie `fra.rêver.1`.

6 Conclusion

By linking heterogeneous information units included in one lexie with other units of other lexies of other languages using first an axie and then all the information possible given by the LFs, we can make closer diverging lexical units of two different languages.

Nevertheless, there are some cases that cannot be solved with this way of dealing with the divergences: when the idioms are free, thus not described by any LF. For example, the French noun “abri” (shelter) meaning that we can be safe into it is not directly translatable in Japanese. We usually find in the dictionaries “hinanjo” but it is a hyponym because it is used only in the case of an earthquake, fire or storm. The part of the meaning not covered by “hinanjo” cannot be directly translatable in Japanese. We then use varied free idioms: “chercher un abri sous un arbre” (to take shelter under a tree) would be translated by “ki no shita ni nigekomu” (to escape under a tree) ; un abri sous la pluie (a shelter under the rain), amayadori suru tokoro (where it is possible to shelter from the rain).

If the idea expressed by a periphrastic idiom in a language is lexicalized in another one, in other words, if there is a lexie with the same meaning in another language, the periphrastic idiom and the equivalent lexie must be described in a way or another in their respective monolingual volume, in order to be linked by an axie.

References

1. Blanc, E.: Une maquette de base lexicale multilingue à pivot lexical: PARAX, In: Lexi-comatique et Dictionnairique, Actes du colloque LTT, Lyon septembre 1995, Actualité scientifique, AUPELF-UREF, Montreal, Canada (1996) 43–58

2. Franckel, J. et al.: Objet, complément, repère, Paris, Langages (1994)
3. Kuroda, K.: (2002) Divergences de Traduction : cas des structures argumentales du japo-nais et du français. Papillon 2002 Seminar, 16-18 July 2002, NII, Tokyo, Japan, (2002)
4. Heid U.(1993) Le lexique : quelques problèmes de description et de représentation lexi-cale pour la traduction automatique, Traductique, AUPELF-UREF, Montréal, Canada, pp.169-196.
5. Lazard, G.: L'actance, Paris, Puf (1994).
6. Lepinette B. (1996) Le rôle de la syntaxe dans la lexicographie bilingue, in Les diction-naires bilingues, AUPELF-UREF, Montréal, Canada, pp 53-66.
7. Levin, B.: English Verb Classes and Alternations, Chicago and London, The University of Chicago Press (1993)
8. Mangeot-Lerebours, M.: Proposal Changes for the Monolingual XML Schema. Papillon 2002 Seminar, 16-18 July 2002, NII, Tokyo, Japan, (2002)
9. Mangeot-Lerebours, M.: Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue. Thèse de nouveau doctorat, Spécialité Infor-matique, Université Joseph Fourier Grenoble I, jeudi 27 septembre 2001, (2001)
10. Mel'čuk I., Clas, A., Polguère, A.: Introduction à la lexicologie explicative et combi-natoire, Paris/Louvain-la-Neuve: Duculot, (1995)
11. Mel'čuk I. (1997) Vers une linguistique Sens-Texte. Leçon inaugurale. Paris: Collège de France
12. Mel'čuk I., et al., (1984, 1988, 1992, 1999) Dictionnnaire explicatif et combinatoire du français contemporain, Vol. I, II, III, IV, Presses de l'Université de Montréal, Mon-tréal, Canada.
13. Montaut, A. et al.: Sur la transitivité dans les langues, LINX, vol.24, Paris, Diffusion Editions Européennes Erasme (1991)
14. Nomura N, Jones, D.A. & Berwick, R.C. An Architecture for a Universal Lexicon COLING-94, 1994, Kyoto", pp 243-249.
15. Polguère, A.: (1998) La théorie Sens-Texte. Dialangue, Vol. 8-9, Université du Québec à Chicoutimi, (1998) 9–30
16. Polguère, A.: Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. Proc. EURALEX'2000, Stuttgart, (2000) 517–527
17. Sérasset, G.: SUBLIM: un Système Universel de Bases Lexicales Multilingues et NADIA: sa spécialisation aux bases lexicales interlingues par acceptions. Thèse de nou-veau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble I (1994)
18. Sérasset, G., Mangeot-Lerebours, M.: Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links. Proc. NLPRS'2001, Hitotsubashi Memorial Hall, Na-tional Center of Sciences, Tokyo, Japan, 27-30 November 2001, vol 1/1, (2001) 119–125
19. Tesnière, L.: Eléments de syntaxe structurale, Paris, Editions Klincksieck (1988)
20. Vandooren, F.: Divergences de traduction et architectures de transfert, Traductique, AUPELF-UREF, Montreal, Canada, (1993) 77–90