MULTILINGUAL LEXICAL DATABASES, LEXICAL GAPS, AND SIM*U*LLDA

Maarten Janssen: *ERSS, Université de Toulouse-Le Mirail (5, allées Antonio Machado, F-31058 Toulouse Cedex 1)*

Abstract

In the design of a Multilingual Lexical Database, one of the biggest problems is constituted by conceptual mismatches between languages, and the resulting matter of lexical gaps. Lexical gaps concern words for which there is no direct translation in a target language, but which nonetheless need to receive a translation within the system. In this article, it will be shown that the various possible ways of dealing with these lexical gaps can be classified in four basic groups. Using the SIM*u*LLDA system as an example (Janssen, 2002), the advantages of the structured interlingua approach over the other possibilities will be explained. With the SIM*u*LLDA set-up, it is possible to derive correct lexical definitions for lexical gaps from the MLLD. How this process of "lexical gap filling" works will be shown using a concrete example of a lexical gap: the treatment of the English words *river* and *stream* in contrast with the French words *fleuve* and *rivière*.

1 Introduction

There is a large number of projects on multilingual lexical databases (henceforth MLLD's). This is not surprising, given the increasing role of computers and the globalisation of the world, creating more contact between languages. One motivation for MLLD's is that it would be desirable to have bilingual dictionaries between all of the approximately 5.000 languages in the world. Since creating over 25 million dictionaries by hand is not a viable option, a more structured approach should be sought.

The set-up of a MLLD would be straightforward if only all languages would have words with the same meanings. But as is generally recognised, for at least two reasons they do not. The first reason is that not all senses of a word have to translate to the same word in the target language. An example is that that the English word *bank* translates to *bank* in Dutch when the financial institution is intended, but to *oever* when the side of a river is meant. This is easily resolved by linking the languages not at the level of their words, but by means of their meanings.

The second reason is the existence of lexical gaps: not every word sense has a direct corresponding word(sense) in every target language (a translational synonym). For instance,

the Russian word *голубой* (goluboj) would be translated as *blue* in English, but *blue* is not a complete translation, since *голубой* is specifically light blue, a colour for which there is no single word in English. In such a case, we say that there is a lexical gap in English for the word *голубой*[1].

Lexical gaps are not omnipresent, but not very rare either: in the MultiWordNet project, a study was done on the Collins English-Italian dictionary, reporting that 5% of the English lexical entries had a lexical gaps in Italian (Bentivogli & Pianto, 2000)[2]. Given this relatively high percentage of lexical gaps, there is a need for a structural treatment of them. In my thesis (Janssen, 2002), a multilingual lexical database is presented, which uses a structured interlingua. This system is called SIM*u*LLDA, and it is capable of handling such lexical gaps. In this article, the general set-up of the SIM*u*LLDA system will be sketched, as well as how such lexical gaps are dealt with in it: not only can lexical gaps be correctly modelled in the system, but it is possible to derive sense-descriptions for bilingual dictionaries from the system. But the next section will be dedicated to a classification of method for dealing with lexical gaps.

2 Dealing with Lexical Gaps

In order to deal with lexical gaps in a proper way, a MLLD needs to somehow 'fill' these gaps; a lexical gap basically concerns a word that has no translational synonym to relate to, but still it needs to be connected to something cross-linguistically. Amongst the many MLLD systems there are many different strategies for doing this, but these strategies can be classified in four major categories. To compare these different strategies, it is useful to consider a concrete example of a lexical gap: the Spanish word *dedo* has no equivalent in English, since English only has the more specific words *finger* (=dedo della mano) and *toe* (= dedo del pie). So in English, there is a lexical gap for *dedo*, and in Spanish there is a lexical gap both for *finger* and for *toe*[3].

There are basically two methods of dealing with these lexical gaps[4]. The first method of 'filling' these gaps is the *project-down* approach, illustrated in figure 1. In the project-down approach, the wordsense of the hyperonymic word *dedo* is 'discarded' and replaced by the two more specific meanings of *finger* and *toe*. The sense distinction between *finger* and *toe* is hence introduced into Spanish, effectively removing the lexical gaps.

The second method is the *hyperonymic* approach, illustrated in figure 2. In the hyperonymic approach, the word *dedo* is explicitly modelled as a hyperonym of the words *finger* and *toe*. This does neither fill nor remove the lexical gap, but acknowledge the existence of a lexical gap, which makes it possible to deal with it correctly afterwards. Both strategies will be evaluated here, and illustrated by showing which approach is used in which way in some existing MLLD systems.

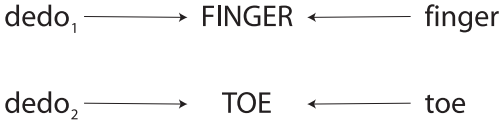

Figure 1: Project-Down

Figure 2: Hyperonymic

Examples of systems that use a project-down approach are Acquilex (Copestake, 1992) and Dhydro (van Campenhoudt, 2001). The set-up of Acquilex is illustrated in figure 3. Acquilex uses feature structures in both the source language (SFS) and the target language (TFS), and at either end at two different levels: 0 for the word level, and 1 for the meaning level. The elements of SFS1 and TFS1 are linked as translatable by *t-links*. The existence of the lexical gap means that in this case, a single SFS1 has to be mapped onto two TFS1's. To solve this, two individual copies of SFS0 (and SFS1) are made.

The set-up of Dhydro is illustrated in figure 4. In Dhydro, lexical gaps are filled in three steps. First, there are three nodes in the interlingual network, which are related, but each of which is expressible only in one of the language. In the second step, the meaning of the hyperonymic term is 'copied onto' the hyponymic nodes (*hyperonomase*). Finally, the now redundant hyperonymic node is discarded (*phagocytée*). As a result, there are two copies of the hyperonymic term, each of which is linked to one of the hyponymic terms.

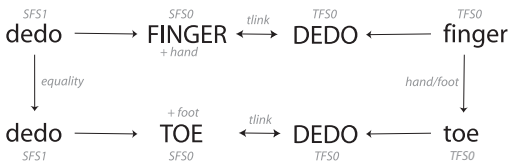

Figure 3: Acquilex (after Copestake, 1992)

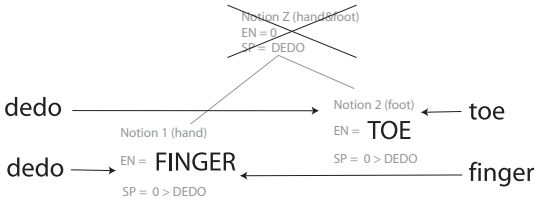

Figure 4: Dhydro (after van Campenhoudt, 2001)

Although the project-down approach effectively solves the problem of lexical gaps, there are at least three fundamental objects against such an approach. The first is a methodological matter. In a way, the project-down approach is methodologically unsound: in the *dedo* example, it introduces an ambiguity in Spanish that is not native to the Spanish language. Even though it might work in practice, the way the meanings are modelled is theoretically not true to the facts. For the Spanish word is not ambivalent between hand an foot, but a more general term. Maybe more appealing: one would not like to say that the word *blue* is ambiguous because Russian distinguishes two variants, nor that *river* is ambiguous because French uses more specific terms.

The second objection is the following: the duplication of meanings that results from projecting-down can lead to an explosive number of meanings in a multilingual setting. Although cases in which this actually happens might be rare, a good illustration of the idea is given by Lyons (1968) – related to this problem by van Campenhoudt (1994: 68). The words for colours are not equally distributed throughout languages, but there are mismatches as illustrated in figure 11. If we want to fill the lexical gaps in this example by means of duplication, as is done in the project-down approach, we need twelve different meanings (indicated by the grey columns) while none of the languages has more than five terms for colours. With more languages, this can rapidly deteriorate.

The third objection is that a database using a project-down set-up is hard to maintain for the following reason: consider a database which already contains Spanish, Italian, Maori, Aramaic, and Swahili, and to which English is added. Spanish, Italian, Maori, Aramaic, and Swahili all have a single word for both fingers and toes. The project-down approach

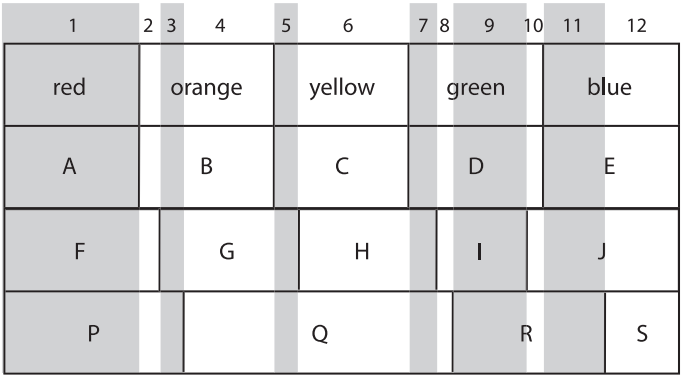

Figure 11. Distribution of Colour Words (after Lyons, 1968)

would require a division of all these meanings, given the distinction between *finger* and *toe* in English. This would mean that all the existing entries for *dedo*, *dito*, *matimati*, *etsba* and *kidole* have to be updated.

All these problems are avoided by the hyperonymic approach. And there is a fundamental reason for that: the word *dedo* is a more general word than either *finger* or *toe*, so it is best to treat it that way. However, the explicit modelling of the fact that *dedo* is not a translational synonym, but rather a translational hyperonym of *finger* forces the issue in what sense *finger* is more specific than *dedo* – what are the differentiating characteristics making *finger* more specific than *dedo*? Unless the differentiae specificae are modelled within the system, there is nothing distinguishing *finger* and *toe* as possible translations for *dedo*.

The hyperonym approach can be subdivided into three variants. The first variant is one without an interlingua (as in figure 6 - also called the *transfer approach*), where the hyperonymy links are present between the language-dependent word-senses. The second variant uses an unstructured interlingua (as in figure 2), and in that case the hyperonymy links are between the words (or word-senses) and the interlingual meanings. And the third variant is a set-up with a structured interlingua (as in figure 7), where the hyperonymy links are between the various meanings in the interlingua.

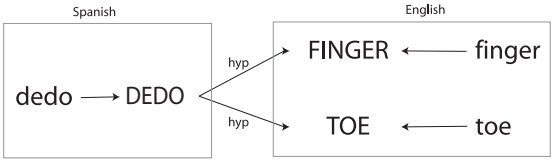

Figure 6: No Interlingua

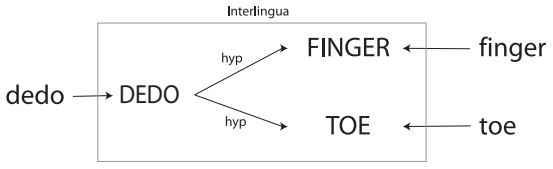

Figure 7: Structured Interlingua

An example of a non-interlingual hyperonymic system is OMBI (Martin & Tamm, 1996) and its multilingual extension in the Hub-and-Spoke model (Beeken *et al.*, 1998)[5], illustrated in figure 8. In OMBI, each language has Lexical Units (LU's) and Form Units (FU's - the meanings). The FU's of the various languages are linked, either as equivalent, or as hyperonym/hyponym. The problem of a non-interlingua set-up such as OMBI is that it is not truly multi-lingual, but more a collection of bilingual connections; all languages are linked in pairs - so by itself OMBI is simply not a MLLD system. In the Hub-and-Spoke model, this problem

is resolved by assigning one of the languages to role of a *hub* to which all other languages can be linked as spokes (hence effectively taking one of the languages as the interlingua)[6].

In the Hub-and-Spoke model every hyperonymy-link is 'decorated' with the appropriate differentiating information. In the case of *finger*, the *hyp*-link has 'della mano' to indicate that a finger is a *dedo della mano*. Although this nicely solves the problem of underspecification, there is a problem with this set-up: the differentiating *della mano* is a string; a language specific element for Spanish. Now if English is taken as a hub, and Swahili is linked to English as well, the word *kidole* will be linked to *finger* as well, with a similar differentiating element: *cha mkono*. The problem is linking *kidole* and *dedo* in the proper way: when they are both linked to *hand*, they will appear as nothing more than translational hyperonyms of the same sense in English, whereas they should be linked as translational synonyms. To be able to arrive at translational synonymy, the two items *della mano* and *cha mkono* now need to be identified somehow. In the Hub-and-Spoke model, these distinguishers are free-text items, and the only way to assure that *della mano* and *cha mkono* indicate the same distinguisher is by explicitly representing their identity.

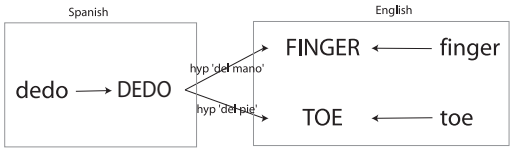

Figure 8: Hub-and-Spoke (after Beeken et al, 1998)

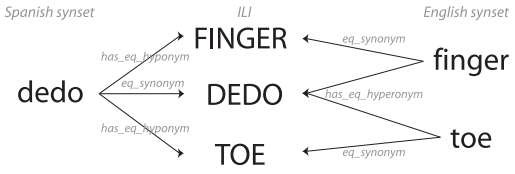

Figure 9: EuroWordNet (after Vossen, 1997)

An example of a hyperonymic method with an unstructured interlingua is EuroWordNet, as illustrated in figure 9. In EuroWordNet, the interlingua consists of an unstructured list of InterLingual Items (ILI's), to which all the synsets of the WordNets of the various languages are linked. There are two problems with this set-up. The first problem is that there are no differentiae specificae in (Euro)WordNet: *finger* and *toe* cannot be distinguish in this set-up. And since the hyperonymy links are not part of the interlingua, but situated between the interlingua and the various languages, it is hard to see how differentiae specificae could be added without getting into the same problem as the hub-and-spoke differentiae have (post-hoc identification).

The second problem is that in EuroWordNet, contrary to the example in figure 7, both the hyperonym and the hyponyms are present in the interlingua. This is in itself an advantage - it avoids the unfortunate property of the set-up in figure 2 that not every language is represented equally in the interlingua. But it has the disadvantage that the word *dedo* needs to be linked to 3 different ILI's. With more languages, the number of links for a single word (synset) can increase rapidly. The reason for this problem is that since the hyperonymy relation is between the languages and the interlingua, the hyperonymy needs to be reestablished for every individual language. An additional risk with this is that the system presupposes a coherent symmetry in the way languages are linked to the interlingua: since *dedo* is equivalent to DEDO and hyperonymous w.r.t. FINGER, and FINGER is equivalent to *finger*, *finger* should also be linked as hyponymous with respect to DEDO. But there is nothing stopping *finger* from being linked as identical to or even a hyperonym of DEDO, which would result in a self-contradictory situation.

The set-up that avoids the problems sketched above is the structured interlingua approach (figure 7). In the structured interlingua approach, there is no need for a duplication of words or meanings, and since the differentiae specificae can be modelled within the interlingua, there will be only one (interlingual) differentiam *of the hand*. Examples of structured interlingua based theories are ULTRA (Farwell *et al.*, 1993) and NADIA (Sérasset, 1994) (see figure 10). None of the current structured interlingua based system does have a way of representing differentiae specificae however. It should in principle be possible to add differentiae to a structured interlingua approach like NADIA in a proper way. However, rather than discussing how this could be done, the next section will present the SIM*u*LLDA system as an example of a structured interlingua approach with differentiae specificae. It could also be argued that knowledge or ontology based systems like KBMT (Nirenburg, 1989) and KRAFT (Visser & Tamma, 1999) also use a structured interlingua. Given their different structure and purpose of these projects, however, they will not be discussed here.

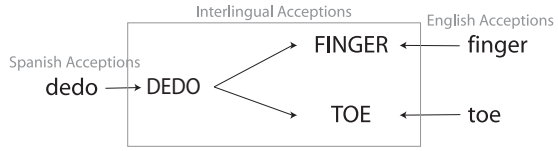

Figure 10: NADIA (after Sérasset, 1994)

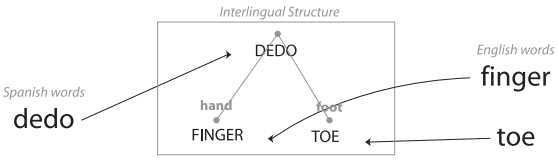

Figure 11: SIM*u*LLDA

The system proposed in my thesis, SIM*u*LLDA, falls in the category of structured interlingua approaches. The set-up of SIM*u*LLDA is illustrated in figure 11. Every word of every language relates to as many interlingual meanings as it has senses (in this example one for each word), and the interlingual meanings themselves are related hierarchically: the meaning DEDO is a hyperonym of both FINGER and TOE, where FINGER is marked as having the additional feature (called definitional attributes in the system) labelled as **hand**, whereas TOE has the definitional attribute **foot**. Because these definitional attributes are part of the interlingual structure, they themselves can be lexicalised in the various languages. So the lexicalisation of **hand** in Spanish is *della mano*, whereas in English it is lexicalised as *of the hand*, and in Swahili as *cha mkono*. In the next section, I will give a more detailed analysis of the set-up of the SIM*u*LLDA system and its virtues.

3 SIM*u*LLDA

In my thesis (Janssen, 2002), a multilingual lexical database system is proposed, which is called SIM*u*LLDA (a Structured Interlingua MultiLingual Lexical Database Application). The aim of SIM*u*LLDA is to provide a tool for lexicographer to aid in the generation of bilingual dictionaries. Since it aims at being a tool, SIM*u*LLDA does not, unlike many other lexical database set-ups, try to criticise or improve the current contents of (bilingual) dictionaries. Rather, the current contents of dictionaries are taken at face value as the starting point for the MLLD.

So the core of the SIM*u*LLDA system consist of dictionary data. The set-up in figure 11 hence more or less directly represents dictionary data, transformed into a structured hierarchy by means of logical tools. A central focus of the thesis is the nature of this logical tool and

the transformation from dictionary data to structure. However, for the topic of the current article, it is only the resulting structure that is of central importance. Nevertheless, here is a very brief sketch of the idea behind the transformation. Since in the SIM*u*LLDA set-up, the exact status of all the various components of the system is very important, some terminology and typography will be introduced in the process.

To illustrate the transformation, a simple example will be treated: the words for horses in English. The relevant definitions are given in table 1 (these are cleaned up versions of the definitions in LDOCE[7].

colt	a young male horse
fil·ly	a young female horse
foal[1]	a young horse
mare	a fully-grown female horse
stal·lion	a fully-grown male horse

Table 1: Definitions of Words for Horses

The definitions in table 1 are analysed in SIM*u*LLDA as relating English words to defining aspects of the meanings expressed by these words. The defining aspects are called *definitional attributes*. So the first definition relates the word colt to the definitional attributes **male** and **young** (as a convention, word-form will be type-set in sans-serif, definitional attributes in **bold face**, and interlingual meanings in SMALL CAPS). On top of these definitional attributes, colt is related to a sense of horse. And the word horse itself is related in the dictionary related to definitional attributes and a further genus proximum, in that case a meaning of the word animal. In this way, lexical definitions can be 'unravelled' into sets of definitional attributes[8]. Thus the dictionary definitions are seen as relating English words and definitional attributes as given in table 2.

	horse	**male**	**female**	**adult**	**young**
HORSE	×				
STALLION	×	×		×	
MARE	×		×	×	
FOAL	×				×
FILLY	×		×		×
COLT	×	×			×

Table 2: Analysis of Definitions for Horses

The rows in table 2 are in fact not words, but meanings of words. And given the interlingual set-up of SIM*u*LLDA, they need to be taken as *interlingual* meanings. So the fifth row of table 2 should be read as indicating that the interlingual meaning FILLY (which is lexicalised in English as filly, and in French as pouliche) 'consists of' three definitional attributes: **horse** (expressed as *horse* in English, and *cheval* in French), *female* (*female* or *femelle* respectively), and **young** (*young* or *jeune*). So this table in fact represent interlingual links and hence not only relates to the English words in table 1, but also to the French words cheval, poulain, pouliche, étalon, and jument (and of course the relevant words of other languages as well).

In the SIM*u*LLDA set-up, the data in table 2 serve as the basis for the interlingual structure. The interlingual structure is in fact no more than a rule-based transformation of these data, the resulting structure of which is given in figure 12.

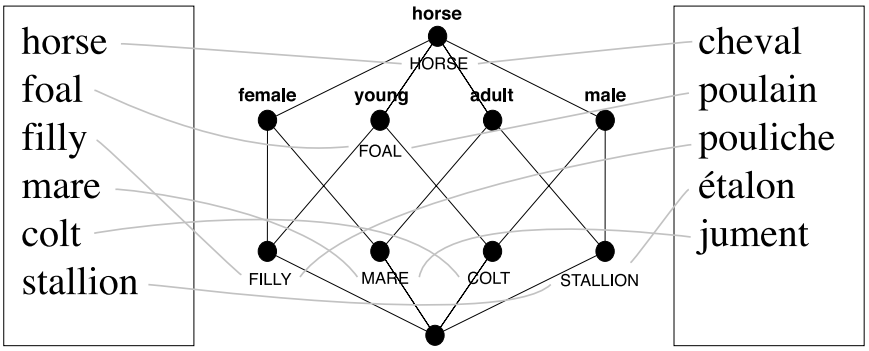

Figure 12. Concept Lattice with Words

The transformation from the table to the structure is done using a logical system called *Formal Concept Analysis* (FCA), developed by Ganter & Wille (1996). It is beyond the scope of this article to explain the workings of FCA, but to give a very quick indication on the basis of figure 12: every node in the structure represents a *formal concept*, and a formal concept is no more than a collection of rows from the table that share crosses in the same rows. For more information about FCA, as well as an on-line tool (called JaLaBA) to perform the transformation, see the web-site of my thesis: `http://maarten.janssenweb.net/simullda`.

Notice that since the data in table 2 were derived from entries in monolingual dictionaries, the interlingual structure in SIM*u*LLDA is in its basis a structured representation of lexicographic data. To properly interpret the structure: all nodes below the node with *female* above it represent interlingual meanings that are characterised by the definitional attribute **female**. And conversely, the interlingual meaning COLT below the node is characterised by all definitional attributes above it.

On the basis of the fact that the interlingual set-up in figure 12 is a structured interlingua based system with an explicit representation of differentiae specificae, it can deal with lexical gap. How this works can be illustrated using the lexical gap present in the figure: the English word colt has no translational synonym in French. To see there is a lexical gap here, one just has to follow the grey line from colt to COLT and see that there is no French word connected to the interlingual meaning. Given the structure of the interlingua, it is possible to generate a definition in French for the word colt despite its lack of a translational synonym. This is done in the following way: the node in the interlingua for COLT has no French word connected to it. But within the interlingua, the node for COLT is connected to a less specific node: the node for FOAL. And FOAL does have a lexicalisation in French: poulain.

So poulain is an approximate translation of colt, but not a complete one: there is a definitional attribute missing. This makes it a translational hyperonym of colt. To find this missing definitional attributes, follow the lattice upwards and collect all definitional attributes above COLT that are not above FOAL. In this case, that is only **male**. Within the set-up, this means that the complete meaning of COLT is FOAL + **male**. The definition in French for colt can now be found by giving the lexicalisation in French for these two components. The lexicalisation

of FOAL in French is poulain, and the lexicalisation of **male** is *mâle*, which means that the complete translation of colt is *poulain mâle*.

Two points should be made here. Firstly, *poulain mâle* is not the only definition that can be created with this method. Also HORSE is a hyperonymic meaning of COLT, lacking two definitional attributes: **male** and **young**. So an alternative would be HORSE + **male** + **young**, or *jeune cheval mâle*. But all definitions generated in this way should properly describe the meaning of *colt* in French.

The second point is that the created definition is a lexicalisation of only interlingual objects: interlingual meanings and definitional attributes. All of these have a lexicalisation in French, but could also be lexicalised in English. Lexicalising them (back) in English would not lead to a translation, but a monolingual definition of colt: *male foal*. The interesting thing is that this is *not* the definition from LDOCE; the definition from LDOCE is the lexicalisation of HORSE + **male** + **young**: *a young male horse*.

3.1 Rivers and Streams

The example in figure 12 is very useful for explaining the general set-up of SIM*u*LLDA. But to show the advantages of its structured interlingua set-up, it is better to compare SIM*u*LLDA with other approaches using a more life-like and regularly discussed example of a lexical gap: the mismatch between the English terms *river* and *stream* on the one hand, and *fleuve* and *rivière* on the other. The reported difference between the two is that the distinction between the tow English notions is their difference in size (rivers are bigger than stream), whereas the distinction in French is that a *fleuve* runs to the sea, but a *rivière* runs to another river. A typical analysis of these data can be found for instance in Sowa (1993) and Sérasset (1994).

fleuve [flœv] n.m. *-fleuve* XII^{e}. lat. *fluvius* **1**$_{\diamond}$ COUR. Grande rivière (remarquable par le nombre de ses affluents, l'importance de son débit, la longeur de son cours); SPÉCIALT lorsqu'elle aboutit à la mer ◇ GEOGR. Cours d'eau (même petit) aboutissant à la mer. (Petit Robert)	**fleuve** Large *rivière* (remarkable by its numbers of affluents, the importance of its debit, or the length of its run); SPECIALISTIC because it ends in the sea GEOGR. stream of water (even small) that ends in the sea.
fleuve [flœv] n.m. (lat. *fluvius*) Cours d'eau qui aboutit à la mer (Larousse)	**fleuve** Stream of water that ends in the sea
fleuve [flœv] n.m. **1** Cour. Grand cours d'eau aux multiples affluents, qui se jette dans la mer ▷ GEOGR Tout cours d'eau qui se jette dans une mer (Hachette)	**fleuve** Big stream of water with multiple affluents, which ends in a sea GEOGR Any stream of water that ends in the sea

Table 3: Definitions of rivière and fleuve

For the analysis, the four relevant words will be interpreted as being defined as indicated in table 4. Four remarks should be made here: firstly, according at least to the Petit Robert and Noailly (1996), a *fleuve* does not really have to end in the sea, it is just a very large stream. Only as a technical term does it specifically relate to streams that end in the sea. But it is not the purpose of this article to question the lexicographic data: if other definitions would be more appropriate, another table and hence another structure would result. Secondly, according to the analysis found in much linguistic work, such as that of Sowa (1993), a *rivière* does not have to be a river, but can also be a smaller stream. This analysis is neither supported by any of the dictionaries, not by corpus evidence[9], although it might be compatble

with the analysis of Noailly (1996). Thirdly, the English word *tributary* is added to the set of data, since it too relates to the (not) ending-in-the-sea. And finally, the definitional attribute **to_sea** is an interlingual item, for which **to_sea** is just an arbitrary label. Less arbitrary are its English and French lexicalisation: *that runs to the sea* and *qui aboutit à la mer* respectively.

	stream	**large**	**to_river**	**to_sea**	**small**
RIVER	×	×			
TRIBUTARY	×		×		
BROOK	×				×
FLEUVE	×	×		×	
RIVIERE	×	×	×		

Table 4: Definitions for Streams of Water

With the set of data in table 4, the interlingual structure with the related words (the lexicalisations of the definitional attributes are left out) is as given in figure 13.

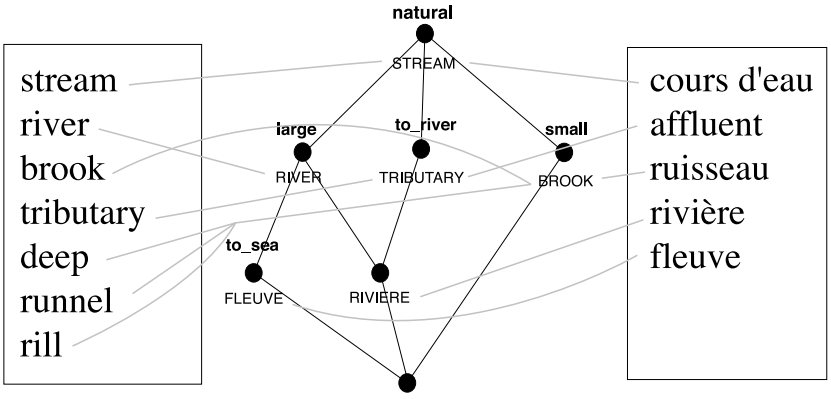

Figure 13. Concept Lattice for Streams of Water

From this structured representation, it is possible to generate bilingual and monolingual definitions for all the relevant words, as was described in the previous section. The result of this process is given in table 5, where the English-French and French-English definitions are given in the top half, and the monolingual definitions in the bottom half. As observed earlier, this table not only contains definitions for words that do have a proper translational synonym, but also for the lexical gaps, such as fleuve.

If we compare this analysis and its results to the analysis of the same data in some other system, the SIMULLDA approach has several advantages where lexical gaps are concerned. Firstly, a comparison with the EuroWordNet (EWN) analysis, which is given in figure 14.

In the EWN approach, words that have a translational synonym in the target language can be translated straightforwardly, and the resulting translation is identical to that rendered by

stream: cours d'eau	**cours d'eau**: stream
brook: ruisseau	**ruisseau**: brook, rill, runnel
tributary: affluent	**affluent**: tributary
river: grand cours d'eau	**rivière**: river that runs to another river
	fleuve: river that runs to the sea
fleuve: grand cours'eau qui aboutit à la mer	**river**: large stream
ruisseau: petit cours d'eau	**brook**: small stream
affluent: cours d'eau qui se jette dans une fleuve	**tributary** stream that runs to another stream
rivière: grand cours d'eau qui se jette dans une fleuve	

Table 5: Definitions Generated by SIM*u*LLDA

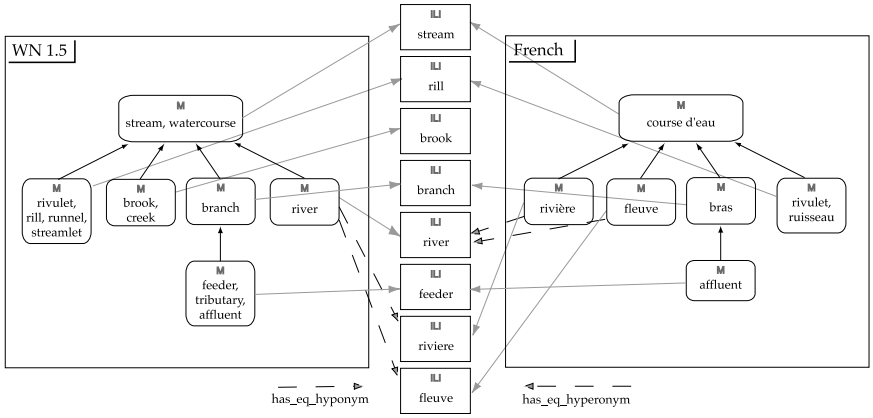

Figure 14. EuroWordNet analysis of Stream of Water

SIM*u*LLDA. As an example: *streamlet* is related to BROOK, and *ruisseau* likewise, therefore *streamlet* and *ruisseau* are translational synonyms.

But where lexical gaps are concerned, the two approaches behave differently. EWN renders the word *river* as the translation of *fleuve*, because there is *has_eq_hyperonym* link from *river* to FLEUVE, and FLEUVE is related with a *has_eq_synonym* link to *fleuve*. Reversely, *fleuve* is linked with a *has_eq_hyponym* link to RIVER, making *river* the translation of *fleuve*. By the same principle, *river* and *rivière* are also linked as translations of each other.

SIM*u*LLDA on the other hand will not render *river* as the translational of *fleuve*. It will link *river* as a translational hyperonym of *fleuve*, but as a translation it will give the more elaborate *river that runs to the sea*. The difference between these two translations is what Zgusta (1971) calls a *translational equivalent* in the case of EWN, and an *explanatory equivalent* in the case of SIM*u*LLDA.

The question which of these two kinds of translations is better is dependent on the purpose: the translational equivalent is more oriented towards the native speaker of the source language for production purposes, and the explanatory equivalent is more informative for the native speaker of the target language for comprehension purposes. Still, there is an advantage of the SIM*u*LLDA approach over the EWN approach here: in many cases, lexical gaps exist because the source language lexicalises a difference the target language does not. And in these cases, the translational hyperonym will be identical to the translational equivalent, as is the case in the *fleuve* and the *colt* example. And whereas translational hyperonyms can be derived from SIM*u*LLDA. explanatory equivalent can never be derived from EWN since

the differentiae specificae are simply missing.

Since the interlingual items themselves are not ordered in EWN, the same hierarchy needs to be represented both between the French WordNet and the ILI's, and the English WordNet and the ILI's. This results in the fact that EWN needs the full Cartesian set of relations between the synsets *river*, *rivière* and *fleuve*, which is clearly redundant from the SIM*u*LLDA perspective.

To look at another system, the analysis by Sowa (1993) of the same set of words is given in figure 15. Sowa in principle uses an interlingual set-up, in which lexical types of different languages are hierarchically linked in a multiple-inheritance structure. And in that respect it has much the same set-up as SIM*u*LLDA.

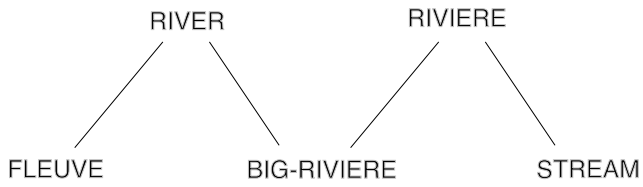

Figure 15. Conceptual Structures analysis of Stream of Water (Sowa, 1993: 246)

But there are two important differences. The first is that in the Sowa set-up, arbitrary lexical types are introduced with the sole reason of relating the various words of different languages. An example of such an artificial lexical type is the item BIG-RIVIERE in figure 15. Neither French nor English has a word expressing this specific meaning. And neither French nor English expresses a hyponym of it. It is simply the intersection of the meaning expressed by *river* and *rivière*. But not the intersection in a technical sense, like the intersection of their definitional attributes as would be the case in SIM*u*LLDA, but the extensional notion of objects being both a river and a rivière at the same time. There are no solid criteria for the introduction of such artificial lexical types, and especially in a multilingual setting, there is a strong risk of a proliferation of such arbitrary items.

The second disadvantage of the Sowa set-up is that is has no (explicit) implementation of differentiae specificae. And the absence of differentiae specificae disallows the creation of explanatory equivalents. That is to say, Sowa explicitly claims that such descriptions should be derivable: "*the word* fleuve *maps into the French lexical type* FLEUVE, *which is a subtype of the English lexical type* RIVER. *Therefore,* river *is the closest one-word approximation to* fleuve*; if more detail is necessary, it could also be translated by the phrase river that runs into the sea.*" (Sowa, 1993: 246). But the problem is that it is by no means clear where the information *that runs into the sea* is supposed to come from. And with the set-up in figure 15 it is also not clear how differentiae could be added in such a way that the structure is coherent: when differentiae are indeed responsible for FLEUVE being hierarchically below RIVER, then a system which explicitly uses them as an ordering principle such as SIM*u*LLDA seems much more natural[10].

4 Conclusion

In this article I have shown the advantages of using a structured interlingua set-up for multilingual lexical databases with an explicit modelling of differentiae specificae. The hyperonymic structure avoids assigning meanings to a language that the language does not express, having the hyperonymic structure in the interlingua avoids having to link languages

pair-wise and avoids having many redundant links. And having differentiae specificae is necessary in a hyperonymic approach to distinguish the various hyponyms of the same hyperonym.

An additional advantage of the structured interlingua set-up is that it allows for the automatic generation of explanatory definitions for lexical gaps, as done by the lexical gap filling procedure in the SIM*u*LLDA set-up. That this is possible is a direct result of the structured interlingua set-up: the structure on the interlingua allows the taxonomic comparison of the non-translationally synonymous terms, and the presence of the differentiae specificae allows to express the difference between the more specific and the less specific terms.

Although not the central topic of this article, I hope to also have indicated that Formal Concept Analysis is a very natural tool for the set-up of a structured interlingua database with differentiae specificae: it is a convenient tool to extract the structure from the relation between the interlingual meanings and the definitional attributes.

Apart from the advantages of the structured interlingua there are of course also some pitfalls. To mention the two most important ones: the risk of an overzealous theory of meaning, and the question of usability. To start with the first: any structured interlingua set-up with differentiae specificae will implicitly, or in the case of SIM*u*LLDA even explicitly, relate interlingual meanings to sets of differentiae specificae. And with such a link, one should be careful not to suggest that concepts can be reduced to limited sets of innate semantic primitives, as was suggested for instance by Katz & Fodor (1963). At least in the case of SIM*u*LLDA, this is not an implication of the system. SIM*u*LLDA is designed to be a lexical database, not a model of mental content: the interlingual meanings in the system are intended to represent only those aspects of word-meaning that are shared cross-linguistically. They are not designed to provide you with the extension of the related words, nor with the associated prototypes, nor resolve any problems regarding the acquisition of concepts. One should be careful not to take the lexical database for more than it is. The problem of the interpretation of the SIM*u*LLDA system is discussed at length in my thesis (Janssen, 2002).

The second problem is the question of usability: in the structured interlingua approach, each interlingual meaning is a hyponym of a more general meaning, where the differentiae specificae are explicitly modelled. This hence presupposes lexical definitions to take the form of genus proximum at differentiae specificae, which is not the case in a great number of examples. There are many definitions in terms of synonyms, meronymic definitions, exemplary lists, etc. In my thesis, it is shown that many of these alternative definitions can be treated within the SIM*u*LLDA system nonetheless. This is done by means of a small empirical study: the treatment of all words for "bodies of water" in six different languages. This study merely focusses on (entity) nouns though. The question whether a structured interlingua approach could be used in practice on a large scale in still an empirical question.

Notes

[1]The notion of a lexical gap is not without problem: one could argue that English does have an expression for this word: *light blue*, which just happens to be a multi-word unit. This would compare to the situation where English uses two words for *computer screen*, while Dutch uses only one (*computerscherm*). Although in my thesis, it is argued that the notion of a 'word' does not relate to a space-separated unit, it is not immediate that there really is a useful notion of a lexical gap. For the purpose of the present article, the existence lexical gaps will be taken for granted, following common practice in many lexical database projects including Acquilex and EuroWordNet.

[2]This number is of course dependent on the definitional gap: in the MultiWordNet count, the word *aniseed* is

considered a lexical gap since its translation (*semi di anice*) is a non-idiomatic multi-word expression.

[3]It could be argued that there is no lexical gap for English here, since either *digit* or *extremity* mean about the same. However, neither of these words is truly a good translation for *dedo*. And even if *dedo* is not really a lexical gap, it should be taken as such for the sake of the argument.

[4]There logically is a third way: simply ignore the difference between *finger* and *toe*. That option does theoretically undesirable, although in practice it is sometimes even applied.

[5]Hub-and-Spoke is a project of the CLVV (Centrum voor Lexicografie en VertaalVoorzieningen, the Dutch/Belgian centre for lexicography and translation) and the IMS (the Institut für Machinelle Sprachverarbeitung of the University of Stuttgart.

[6]In principle, there does not need to be only one hub in the Hub-and-Spoke model: various interconnected hubs can exist like in a computer ethernet network.

[7]Longman Dictionary of Contemporary English, second edition, 1987.

[8]Provided that this process terminates somewhere. In my thesis, the termination is discussed in detail, but for simplicity, I will ignore the genus term horse in this example, and treat **horse** as if it named just another definitional attribute.

[9]In all aligned corpora I have looked at, both *fleuve* and *rivière* are consistently translated as *river*, with one exception: the French phrase *dans fleuves and rivières* occurring twice was translated in both cases by *rivers and their tributaries*.

[10]Notice furthermore that Sowa seems to explicitly state here that translational hyperonyms are the best one-word approximations.

References

Beeken, Jeannine; Heid, Ulrich; Laureys, Godelieve; Martin, Willy, and Schuurman, Ineke. 1998. *On the Construction of Bilingual Dictionaries: feasibility study carried out by order of the European Commission DG XIII*. Technical Report. Stuttgart.

Bentivogli, Luisa, and Pianto, Emanuele. 2000. Looking for Lexical Gaps. *In: Proceedings of the Ninth Euralex International Congress*.

Campenhoudt, Marc van. 1994. *Un Appart du Monde Maritime à la Terminologie Notionelle Multilingue: étude du dictionnaire du capitaine Heinrich Paasch* de la quille à la pomme de mât. Ph.D. thesis, Université de Paris XIII, Paris.

Campenhoudt, Marc van. 2001. Pour une Approache Sémantique du Terme et de ses Équivalents. *International Journal of Lexicography*, vol. 14:181 – 209.

Copestake, Ann. 1992. *The Representation of Lexical Semantic Information*. Ph.D. thesis, University of Sussex, Sussex.

Farwell, David; Guthrie, Louise, and Wilks, Yorick. 1993. Automatically creating lexical entries for ULTRA, a multi-lingual MT system. *Journal of Machine Translation*, vol. 8:127 – 146.

Ganter, Bernhard, and Wille, Rudolf. 1996. *Formale Begriffsanalyse: mathematische grundlagen*. Berlin: Springer Verlag.

Janssen, Maarten. 2002. *SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua*. Ph.D. thesis, Universiteit Utrecht, Utrecht.

Katz, Jerrold J., and Fodor, Jerry A. 1963. The Structure of a Semantic Theory. *Language*, vol. 39:170 – 210.

Lyons, John. 1968. *An Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Martin, Willy, and Tamm, Anne. 1996. OMBI: Aan Editor for Constructing Reversible Lexical Databases. *In:* M. Gellerstamm (ed.), *Proceedings of the Seventh Euralex International Congress*.

Nirenburg, Sergei. 1989. Knowledge-Based Machine Translation. *Machine Translation*, vol. 5:5 – 24.

Noailly, Michèle. 1996. Dans le Sens du *Fleuve*: Syntaxe et Polysémie. *In:* Kadyallah Fall, Jean-Marcel éard, and Paul Siblot (eds.), *Polysémie et Construction du Sens*. Montpellier: Praxiling, Presses de l'Université Paul-Valéry.

Sérasset, Gilles. 1994. SUBLIM *un Système de Bases Lexicales Multilingues et* NADIA *sa Spécialisation aux Bases Lexicales Interlingue par Acceptions*. Ph.D. thesis, Université Joseph Fourier, Grenoble.

Sowa, J.F. 1993. Lexical Structures and Conceptual Structures. *In:* James Pustejovsky (ed.), *Semantics and the Lexicon*. Dordrecht: Kluwer.

Visser, Pepijn R.S., and Tamma, Valentina A.M. 1999. An Experience with Ontology-Based Agent Clustering. *In:* Benjamins, Chandrasekaran, Gomez-Perez, Guarino, and Uschold (eds.), *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5)*.

Vossen, Piek. 1997. EuroWordNet: a multilingual database for information retrieval. *In: DELOS workshop on Cross-language Information Retrieval*.

Zgusta, Ladislav. 1971. *Manual of Lexicography*. Den Haag: Mouton.