Lexical Translation and Conceptual Hierarchies

Maarten Janssen

Abstract
In my thesis a multilingual lexical database is proposed, called SIM*u*LLDA, in which interlingual meanings are organised into a conceptual hierarchy by means of a logical formalism called *Formal Concept Analysis*. The resulting structure is a lattice in which the nodes are organised by means of their attributes, which are abstract representations of the differentiae specificae in dictionaries. This lattice order allows amongst others a proper treatment of *lexical gaps*: words without a translational synonym. But although the lattice ordering solves and clarifies several lexicographic problems, lexicographic practice in some cases demands a more liberal structure in which concepts between which there is not strict relation can be related nonetheless, going against the logical ordering. This article sketches the set-up SIM*u*LLDA set-up, and the conflicting interests of lexicographic practice.

1 Introduction

This paper will illustrate a conflict between a logical approaches to multilingual lexical database and and the demands of lexicographic practice. The logical approach that will be used is called SIM*u*LLDA and is based on the application of a logical framework called *Formal Concept Analysis* to a multilingual lexical database. The resulting system allows cross-linguistic comparison of lexical meanings, which in turn allows a proper treatment of so-called *lexical gaps*: words in one language for which there is no translational synonym in another.

The purpose of the SIM*u*LLDA system is to provide a multilingual lexical database in which every language is linked to a structured interlingua once, and in which translation relations between two languages are derived from logic entailments. However, lexicographic practice shows a desire to have correspondences between words and meanings where logically speaking no strict relation exists. This need arises in the cases of partially overlapping lexical gaps: pairs of words in languages that express intuitively similar, but logically different pairs of meanings.

In this paper, I will briefly sketch the set-up of the SIM*u*LLDA system, and how it allows a multilingual database to deal with lexical gaps. After that, I will discuss the lexicographer's problem concerning overlapping meanings, and the problem it poses for the logical framework.

The basic idea behind the SIM*u*LLDA framework is this: in (monolingual) dictionaries, nouns are generally defined in terms of *genus proximum et differentiae specificae*. That is to say, a specific word-meaning is claimed to be subordinate to another word-meaning, differentiated from it by certain semantic features. These semantic features are called *definitional attributes* within the SIM*u*LLDA system. When combined with the notion of inheritance, this results in a system in which word-meanings are related to sets of definitional attributes, being the recursive collection of all the differentiae specificae of all genus terms.

The resulting sets of definitional attributes can be interpreted as defining a formal context in the sense of Formal Concept Analysis. FCA is a formal attempt to define the notion of a concept within the boundaries of model theory [1]. An FCA context consists of a set of objects G, a set of attributes M, and a relation I relating the two sets, where $(a, b) \in I$ means that object a has attribute b. Within such a formal context, formal concepts are defined as those pairs of formal objects and formal attributes that mutually define each other in the sense that no other objects share all these attributes, and all the objects share exactly that set of attributes. The formal definition of the formal concepts $\mathfrak{B}$ is given belows, and $\langle \mathfrak{B}, \leq \rangle$ is a complete lattice:

$$B^{\downarrow} = \{g \in G \mid \forall b \in B.\ (g, b) \in I\} \tag{1}$$

$$A^{\uparrow} = \{m \in M \mid \forall a \in A.\ (a, m) \in I\} \tag{2}$$

$$\mathfrak{B}(G, M, I) = \{\langle A, B \rangle \mid A = B^{\downarrow} \wedge B = A^{\uparrow}\} \tag{3}$$

$$\langle A_1, B_1 \rangle \leq \langle A_1, B_1 \rangle \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_2 \subseteq B_1 \tag{4}$$

In the SIM*u*LLDA framework, FCA is applied to lexicographic data by taking as formal objects (interlingual) word meanings, and as formal attributes definitional attributes. It takes the tabular structure resulting of the analysis of the genus et differentiae data in dictionaries, and yields a lattice structure. An example of a SIM*u*LLDA context is given in table 1. As a convention, word form are written *slanted*, interlingual meanings in SMALL CAPS, and definitional attributes in **bold face**.

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

Table 1: SIM*u*LLDA Context for Horse Words

For the transformation of a tabular structure to a lattice Hasse-Diagram, an HTML-based tool was created, called *JaLaBA*, which can be found on the web-site of my thesis: `http://maarten.janssenweb.net/simullda`. The JaLaBA applet generates a lattice from this tabular set of definitions, which is given in the middle of figure 1.

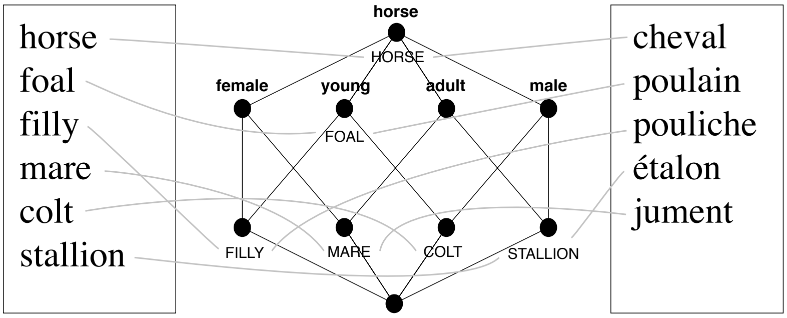

Figure 1: Concept Lattice for Horses

Within the SIM*u*LLDA framework, definitional attributes are taken to be interlingual. That is to say: the differentiam *jeune* used in French dictionaries and the differentiam *young* in English ones are taken to be lexicalisations in different languages of one and the same definitional attribute **young**, which itself is hence language independent. By doing this, the lattice structure resulting from the FCA analysis is an interlingual structure, allowing the comparison of lexical meanings across the various languages that are linked to the interlingua. In the case of figure 1, the word *horse* and the word *cheval* are translational equivalents because they relate to the same interlingual meaning in the interlingua. For a more detailed description of SIM*u*LLDA and its use of FCA, see Janssen [2, 3].

3 Lexical Gaps

A key problem in multilingual lexical database design is the treatment of *lexical gaps*: notions for which there is a word in the source language, but not in the target language. In a paper for the *International Journal of Lexicography*, Marc van Campenhoudt [8] sketches three distinct types of lexical gaps: hyperonymy cases, complex hyperonymy cases, and partial overlapping cases. In this section, I will show that the first two cases are easily dealt with in the SIM*u*LLDA system, but that the third class poses a fundamental problem for SIM*u*LLDA, and in fact any framework trying to deal with lexical gaps in a formal, taxonomy based approach[1].

[1] For an explanation of how van Campenhoudt himself deals with these cases of lexical gaps, see van Campenhoudt [4, 8].

Hyperonymic lexical gaps are cases where there is no direct translational synonym for a word, but there is a translation for the genus proximus. A example of a hyperonymic case of a lexical gap can be found in the example in figure 1: for the English word *colt* there is no direct equivalent in French. In the SIM*u*LLDA set-up, this lack of a translational synonym is given by the fact that the interlingual meaning expressed by *colt* (indicated as COLT) has no lexicalisation attached to it in the French language. But there is lexicalised hyperonym of COLT in French: *poulain*.

Hyperonymic lexical gaps can be dealt with very elegantly in SIM*u*LLDA: any interlingual meaning equals its genus proximus plus the differentiating definitional attributes. In the case of COLT: COLT = POULAIN+**male**. By taking the lexicalisation in English and French respectively of all parts of this equation, we arrive at an explanatory equivalent in the target language – in this case the definition *poulain mâle* for the lexical gap *colt*. So the definition in a target language for a lexical gap is the lexicalisation of the first super-concept for which a lexicalisation in the target language exists, together with the lexicalisation of all the elements of the difference between the set of definitional attributes of that super-concept and the interlingual meaning of the lexical gap.

This method of filling lexical gaps works equally well for the second type of lexical gaps: the complex cases of hyperonymic lexical gaps. Complex hyperonymy cases are cases where the the lexical gap spans more than one taxonomic level. The example van Campenhoudt [8, 3] gives is the English word *plunging breaker*, which does not have a translation in French. Nor does the genus term *breaker* have a translation in French. The translation in French should be *vague déferlante*, which is a more general term than *breaker*.

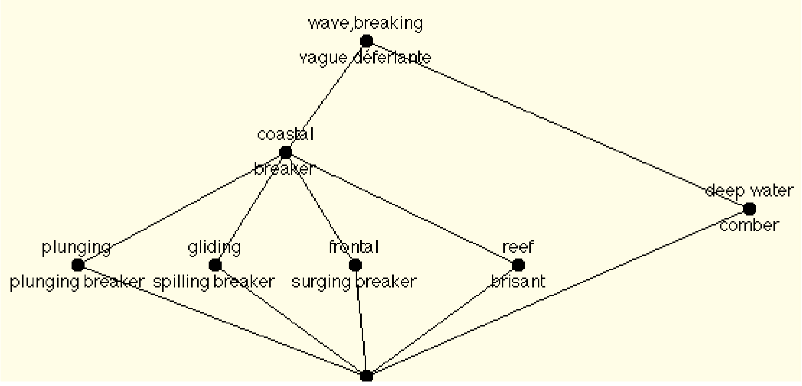

Figure 2: Complex Hyperonymy Case

The SIM*u*LLDA lattice of the lexical field of breaking waves, generated by JaLaBA,is given in figure 2. From the SIM*u*LLDA perspective, the only difference between the simple and the complex hyperonymy case is that there is more than one definitional attribute missing from *vague déferlante* wrt *punging* *breaker*. The general method work the exactly the same: PLUNGING_BREAKER = VAGUE_DEFERLANTE + **plunging** + **coastal**. Lexicalising these different elements in French renders the French explanatory equivalent for *plunging breaker*: *vague déferlante plongeant à la rivage*.

3.1 Overlapping Meanings

The case of *partially overlapping* lexical gaps, contrary to the two other types, poses a serious problem for the SIM*u*LLDA approach. Partial overlaps are those cases in which the meanings of two (pairs of) words of different langauges are intuitively similar, but between the meaning of which no a priori inclusion relation exists. A clear, often cited and intuitive example of partial overlap is given by Sowa [6]: the relation between the Frech words *fleuve* and *rivère* on the one hand, and the English words *river* and *stream* on the other. The cited difference between these terms is this: a *fleuve* ends in the sea, and a *rivière* in another fleuve or rivière, whereas the difference between a river and a stream is just one in size[2].

There is a considerable overlap between the notions *river* and *rivière*, which is more than just an accidental existential overlap: there is the conceptual correlation that larger streams of water will in general end in the sea, whereas the smaller ones are tributaries of the larger ones. However, this is a necessary entailment in neither way, making that there is no strict inclusion in either direction.

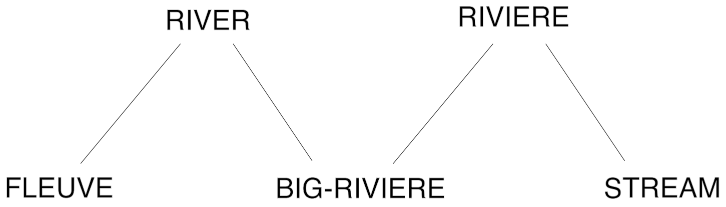

Figure 3: Partially overlapping gaps

The desire of lexicographic practice is to express this large overlap within the hierarchy. In his article, Sowa [6] does this by calling into existing a notion Big-Rivière (see figure 3): *"River ... has two subtypes: one is Fleuve, which maps to* fleuve*, the other is the English-French hybrid BigRivière, whose closest approximation in French is the single word* rivière *or the phrase* grande rivière.*"* [7, p. 410].

[2] This correspondence cited by Sowa is at least partly contradicted by lexicographic evidence, and is also in contradiction with corpus data. In all occurrences in the aligned corpora we queried, *fleuve* and *rivière* appear as hyponyms of *river*, which makes it a case easy to solve within the SIM*u*LLDA framework. But in the situation sketched by Sowa, a rivière can be either a river or a stream depending on its size, and a river can be either a fleuve or a rivière depending on where it ends.

Although intuitively clear, the problem with this solution is that it is logically ill founded. Creating a common hyponym for Rivière and River only indicates that the meanings of Rivière and River are not incompatible. If the taxonomy is to be read extionionally (which in the case of SIM*u*LLDA it is not), it would even mean that there are rivers that are also rivières, but it no more brings the terms closer together than it would make the terms *Japanese* and *woman* translations of each other by introducing a common hyponym JapaneseWoman. Without a mechanism that assures translation in a situation like in figure 3, a common hyponym does not lead to filling the lexical gap.

A possible solution to this is the use of a principle which van Campenhoudt [8] calls *hyperonomase*: in the case of a lexical gap, the more general term is copied onto its (tranlational) hyponyms to allow translation. An example is given in figure 4: the words *bras de mer* and *reach* are partially overlapping[3]. By applying hyperonomase to *reach*, two meanings of reach are created, one which is a translation of *section rectiligne*, and the other which has no translational equivalent in French, i.e. a new lexical gap is created. This newly created lexical gap is then filled by applying hyperonomase to *bras de mer*: *"une double hyperonomase conduit à créer une entrée qui permet de désigner le sémème correspondant a cette intersection partielle."*[4]

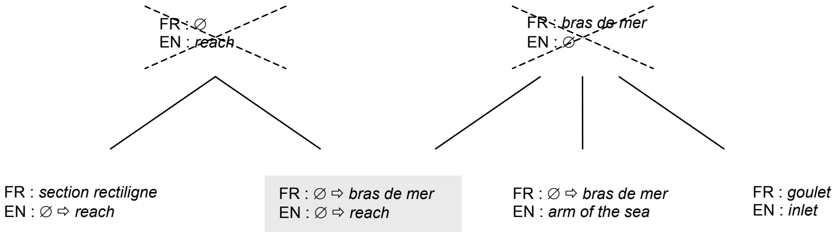

Figure 4: Hyperonomase

Apart from two conceptual problems[5], this solution has a false air of generality: from the graph in figure 4, it appears as if the mere existence of the two lexical gaps for *reach* and *bras de mer* leads automatically to a translation between the two terms. But in fact there is no such immediacy: the only reason to divide the meaning of *reach* into the two distinctions it has been given in figure 4 is to allow linking it to *bras de mer*, because *bras de mer* is the intended

[3] This is seen as a lexical gap only because of the the strict monosemic approach taken by van Campenhoudt: in the SIM*u*LLDA system, *reach* would be considered polysemous from the start, with two normal translational synonyms in French.

[4] A double hyperonomase creates an entry which permits to indicate the meaning corresponding to that partial intersection.

[5] The fact that this solution introduces a distinction in English that the English language does not lexicalise and that it leads to a proliferation of meanings in a multilingual setting, see [4].

translation. The solution basically consists of creating a direct translation relation between *reach* and *bras de mer*, without really using the taxonomy or the componential analysis.

This is more easy made clear with the example BigRivière: the reason for creating a meaning of *river* that involves its ending in the sea or not is to link *river* to the words *fleuve* and *rivière*. There is no real motivation a priori for this linking; it would only be motivated if the reason for their being translation was part of the process: if the linking itself was a result of the correlation between ending in sea/river and size. However, the linkage does not exploit this correlation (strictly speaking it also could not exploit it, since there is no logical entailment in either direction).

The fundamental problem is that in the cases of partially overlapping lexical gaps, there is a mismatch between the taxonomic demarcation of the meaning of the terms, and the question of translatability: what counts as the best translation of the lexical gap is a separate question from what the respective terms mean, and how they are taxonomically related. This is most clearly illustrated by the way the quoted example of *river* and *rivière* would be treated in SIM*u*LLDA: the SIM*u*LLDA lattice would have all four terms as direct hyponyms of *stream of water*, with two distinct sets of differentiae specificae. And to translate in either direction, *stream of water* would have to be the genus term used in the lexical gap filling procedure: for the word *river* SIM*u*LLDA would produce something like *grande course d'eau naturelle* (depending on the actual definitional attributes used), and the other way around, for *fleuve* it would have *natural stream of water ending in the sea*.

Strictly speaking these definitions reflect the only formal relation between the two terms: they are separate divisions of a common genus term. The fact that for all practical purposes, *fleuve* would have to be translated by *river* would be the result of considerations outside of the scope of the decompositional taxonomic analysis[6].

4 Conclusion

Lexical gaps can be very elegantly treated in a taxonomically based decompositional multilingual lexical database system such as SIM*u*LLDA. The hyponymy relations, taken together with the definitional attributes lead to an explanatory equivalent for hyponymic lexical gaps, whether they are simple cases, or complex mismatches spanning several taxonomic levels.

But the treatment of partially overlapping lexical gaps such a system does not lead to the translational relations desired by lexicographic practice. The fundamental reason for that is that the translations desired by lexicographic

[6] In fact, the SIM*u*LLDA example of these data is different, since as mentioned before, the quoted example does not match the lexicographic data: in the SIM*u*LLDA analysis, both FLEUVE and RIVIERE are subconcepts of RIVER, see [2, p. 134].

practice are not the result of taxonomic considerations. The SIM*u*LLDA system hence considers all lexical gaps as hyperononymic lexical gaps, whether simple or complex.

References

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

[2] Maarten Janssen. *SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua.* PhD thesis, Universiteit Utrecht, Utrecht, 2002.

[3] Maarten Janssen. Multilingual lexical databases, lexical gaps, and simullda. To appear in International Journal of Lexicography.

[4] Maarten Janssen and Marc van Campenhoudt. Terminologie traductive et representation des connaissances: l'usage des relations hyponymiques. To appear in Langages.

[5] Igor Mel'čuk and Leo Wanner. Towards a lexicographic approach to lexical transfer machine translation: illustrated by the german-russian language pair. *Machine Translation*, vol. 16:21 – 68, 2001.

[6] J.F. Sowa. Lexical structures and conceptual structures. In James Pustejovsky, editor, *Semantics and the Lexicon.* Kluwer, Dordrecht, 1993.

[7] John F. Sowa. *Knowledge Representation: logical, philosophical and computational foundations.* Brooks/Cole, Pacific Grove, 2000.

[8] Marc van Campenhoudt. Pour une approache sémantique du terme et de ses Équivalents. *International Journal of Lexicography*, vol. 14:181 – 209, 2001.