

Chapter 6

Conclusion and Afterthoughts

As outlined in the first chapter, the purpose of this thesis was a practical one: to construct a system (*SIMuLLDA*), which is a multilingual lexical database that can contain an arbitrary number of languages, and which aims at the following:

1. Bilingual dictionaries between arbitrary pairs of languages from the database should be generated by the system.
2. The system should be a tool for lexicographers, and hence take dictionary definitions seriously (and as much as possible at face value)
3. It should even generate definitions in case the target language has no direct translation of the word in question.

Let me here once again briefly sketch the basic set-up of the *SIMuLLDA* system, and describe whether it meets the requirements formulated above.

The basic lay-out of the *SIMuLLDA* system is illustrated in figure 6.1: for every language in the system there is a language module. These language modules consist of lists of lexemes, and lexemes in turn are sets of word-forms, represented by their citation-form. The role of each word-form in the lexeme is indicated by means of a lexical function, which functionally determines the relation between the citation-form and the word-form in question.

All the language-modules are related to the interlingua. The structure of this interlingual is the heart of the *SIMuLLDA* set-up. The interlingua consists of a lattice structure, the nodes of which are pairs of interlingual meanings and definitional attributes.

The interlingual meanings are the senses expressed by the lexemes from the various languages. Every lexeme in every language expresses one or more interlingual meanings. But that does not mean that every interlingual

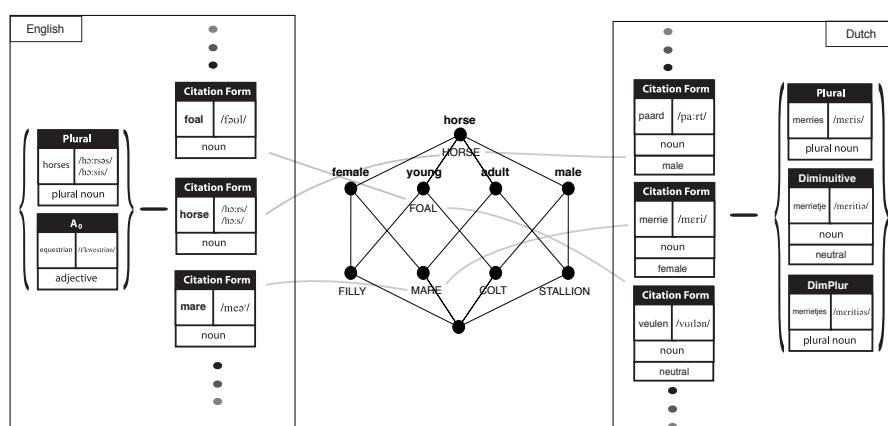


Figure 6.1: Set-Up of the SIMuLLDA System

meaning has to be expressed in every language: if Italian has a specific word for roads that lead to Rome, there will be an interlingual meaning that is only lexicalised in Italian. So there can be lexical gaps in the system.

The definitional attributes are the properties that define the interlingual meanings. Like the interlingual meanings, the definitional attributes are linked to all the various languages, as is illustrated in figure 6.2. Definitional attributes are not expressed by means of word-forms or lexemes, but by means of strings. So in figure 6.2, the definitional attribute **young** is expressed by *jeune* in French.

The nodes of the interlingual concept lattice are formal concepts: pairs of interlingual meanings that share a set of definitional attributes, combined with the definitional attributes that they share. In figure 6.1, the definitional attributes are represented above the highest node in which they appear, and the interlingual meanings are represented under the lowest node in which they appear. So in the figure, FOAL, FILLY, and COLT all have the definitional attribute **young**, since **young** is higher than all of them. And FILLY has all the attributes **young**, **female**, and **horse**, since FILLY is under all of them.

In figure 6.1, every node in the lattice represents a formal concept, and hence has a number of definitional attributes related to it. By definition, every higher node that is connected to that node has a subset of these definitional attributes. Therefore, the lower nodes have a surplus of definitional attributes over the higher ones. And this definitional surplus allows us to generate definitions for words that do not have a direct translation, i.e. for lexical gaps.

A lexical gap in SIMuLLDA is defined as a lexeme x of some source language, that relates to an interlingual meaning for which there is no related lexeme in another language Y . In such cases then say that there is a lexical

gap in Y for x .

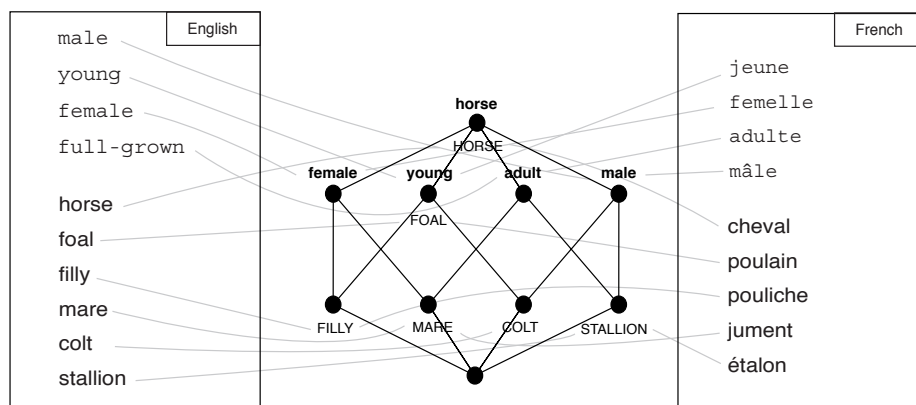


Figure 6.2: Lexical Gap in SIMuLLDA

To take an example: the English word *colt* has no translation in French. There is a word for young horses (*poulain*), and a word for female young horses (*pouliche*), but not for male young horses. This situation is illustrated in figure 6.2.

Normally, the translation of a lexeme is found by going from the citation-form to the interlingual meaning it expresses, and then to the word-form that is related to that same interlingual meaning in the desired target language. So the word *foal* relates to the interlingual meanings *FOAL*, which is expressed by the word *poulain* in French. This results in saying that the translation is found by ‘following the lines’ in figure 6.2.

Since there is a lexical gap in French for the word *colt*, following the lines does not work: the English word *colt* relates to *COLT*, but there is no French lexeme related to that interlingual meaning. Such lexical gaps can be ‘filled’ in SIMuLLDA by the lexical gap filling procedure (section 2.3.2). The way this works is as follows: the interlingual meaning *COLT* has no French word related to it. But the node for *COLT* is a subnode of the node for *FOAL*, and for *FOAL* there is a French lexicalisation too: *poulain*. There is a difference between *COLT* and *FOAL*: since the first is a subnode of the second, *COLT* has a definitional surplus over *FOAL*. This definitional surplus is **male**. For **male**, there is a lexicalisation in French: *mâle*. So to get the complete meaning of *colt* in French, these two lexicalisations have to be taken together: *poulain* as genus proximum, and *mâle* as the differentiam specificam. And *poulain mâle* is indeed the (explanatory) translation of *colt*.

The interlingual lattice in figure 6.1 is not entered as such in the SIMuLLDA set-up, but a result of a logical system (Formal Concept Analysis) that brings structure to the unstructured data underlying the lattice. These unstructured data consist of cross-tables, in which the rows are interlingual meanings, and the columns are definitional attributes. FCA builds a lattice out of the cross-table by defining the nodes to be all pairs of interlingual meanings and definitional attributes for which it holds that all meanings have all attributes and vice versa, and defining the order on the nodes by the subset relation on the sets of attributes (see chapter 2).

The set-up in figure 6.1 is related directly to dictionary definitions in two ways: on the one hand, monolingual English dictionary definitions can be derived from the structure in figure 6.1, and on the other hand, the structure in figure 6.1 can be derived from the relevant definitions in a monolingual English dictionary. Deriving definitions from the structure can be done by taking the lexical gap filling procedure, and translate from source language to source language. This will precisely yield the dictionary definitions for all the lexemes in the language: for stallion it will give *male horse*, for mare it will give *female horse*, etc.

Deriving the structure from dictionary definitions can be done by taking the dictionary definition of say *colt* (which is *male young horse*), take the two differentiae specificae in this definition as expressing definitional attributes (**male** and **young** respectively), and take these, together with the definitional attributes of the genus term, as the definitional attributes defining the interlingual meaning. If we do this for all the words for horses in English, we get the kind of cross-table that can be structured by FCA, resulting in the concept lattice in figure 6.2.

In chapter 4, it was shown that this converting dictionary data to SIMuLLDA concept lattices can be done at large scale, and that we get the appropriate bilingual definitions from them. This was done for the words for horses in Italian, English, and Russian; for the words for bodies of water of six languages (English, Dutch, German, French, Italian, and Russian), and for the words for the sails on a ship in English, German and French.

Because of this double dependency between the SIMuLLDA set-up and dictionary data, we can state that the SIMuLLDA set-up models dictionary data retrievably, and hence can function as a lexicographers tool. And since, as described above, it is a lexical database that can generate bilingual definition (for every pair of language), even in case of a lexical gap, the SIMuLLDA set-up meets the three requirements this thesis aimed at.

6.1 Program for Further Research

As observed in chapter 5, the system presented in this thesis is not a complete multilingual lexical database application. Some of the lacking features were discussed there, with proposals to their solution in *SIMuLLDA*. But these proposals all leave a number of open questions.

One such open question was discussed in the previous chapter (section 5.3): how well does the *SIMuLLDA* system come out when applied to other word-classes, such as verbs? The discussion of this question will be even more complicated than that of terms for bodies of water in chapter 4, for two reasons: since verbs are less nicely grouped into lexical fields, it is even harder to find all the appropriate lexical entries, and the interlingual alignment of verbs is even harder than that of entity nouns, since they are often even harder to define, and their differences are also often intertwined with their differences in grammatical behaviour.

Another open question is whether the proposed use of lexical functions for the modelling of derivations really works at a large scale, and which lexical functions actually play a role in this. And whether the proposed way of restricting the output of the system by means of saliency indices actually works in practice, and what the optimal number of indices is. Furthermore, it could be looked into whether saliency should also play a role in the definition generation: should we allow words with a very low saliency index to appear as translational synonyms in a small dictionary?

And then there are some open problems apart from those related to the extensions in chapter 5: by its formal design, the *SIMuLLDA* system puts some restrictions on the way in which lexical definitions can be given. This was avoided as much as possible, since *SIMuLLDA* is designed to be a lexicographic tool rather than a lexicographer's annoyance. But still, the formal set-up forces a more explicit way of designing lexical entries. The most pressing restriction on dictionary definitions is the fact that in all circumstances, an existing meaning for the genus term in the lexical definition has to be selected. And as discussed before, this can be difficult in some situations, for instance in the case of hyponyms of regular polysemes. The question is whether this requirement would be too restricting for lexicographic practice. But it is also clear that without such a restriction, dictionaries are too informal to be captured in a formal system.

Lexical gaps in the *SIMuLLDA* set-up are filled by the lexical gap filling procedure with a description in terms of *genus proximum et differentiae specificae*. And as we have seen, this description is in most cases the direct translation of the monolingual definition of the word in the source language. So if a lexical entry in the source language has a lexical gap in the target language, its definition will be (most often) the monolingual dictionary entry of the source language, literally translated into the target language. This is in principle a valid method, that should yield proper

translations. But when looking at actual dictionary definitions, there often is a structural difference between the explanations given in the bilingual dictionary in case of a lexical gap, and the definition given in the monolingual dictionary. These differences mainly concern the level of explanation and the choice of genus terms. For a solid analysis of the SIMuLLDA system, these differences should be further specified and the question should be answered whether these are desirable differences, or more the product of lexicographic tradition. When these differences are necessary, the translations yielded by SIMuLLDA might not be as useful as you might hope. No such analysis was given in this thesis, for a proper analysis would require a large amount of lexicographic data, and would be more fit for an empirical setting than for a theoretic analysis as the one given in this thesis.

In the light of all these points of further research, a lot of work has to be done to reach the pragmatic goal set out by this thesis: to have a multilingual lexical database from which complete bilingual dictionaries can be generated. And the only way to really discover if it works in practice, and find an answer to the open questions above is to have a full implementation of the system, and fill it with lexicographic data.

But in this thesis I hope to have shown that the SIMuLLDA set-up provides a useful framework for a multilingual lexical database, and that FCA is a useful tool for the multilingual alignment of lexicographic data.