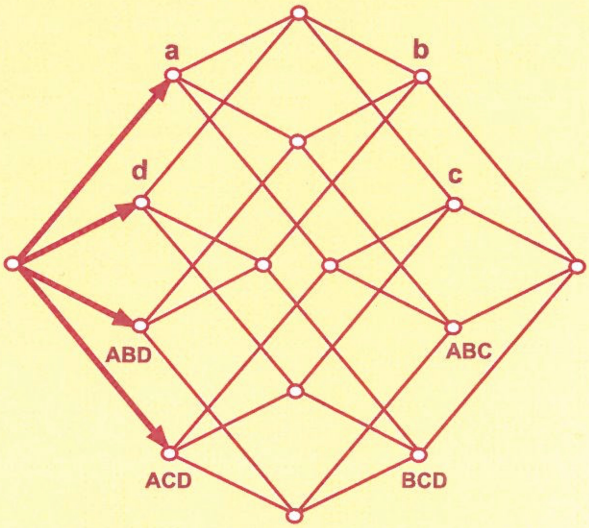
SIM*u*LLDA

a Multilingual Lexical Database Application
using a Structured Interlingua



PhD Thesis Cognitive Artificial Intelligence
Maarten Janssen, June 7, 2002

SIM*u*LLDA
a Multilingual Lexical Database Application using a Structured Interlingua

SIM*u*LLDA
een toepassing van een meertalig lexicaal gegevensbestand
met gebruikmaking van een gestructureerde tussentaal
(met een samenvatting in het Nederlands)

Proefschrift
ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op het gezag van de Rector Magnificus, Prof. dr. W.H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 7 juni 2002
des middags te 4:15 uur

door
Maarten Janssen
geboren op 28 januari 1971 te Nijmegen

Promotoren:
Prof. dr. H.J. Verkuyl
UiL-OTS, Universiteit Utrecht
Prof. dr. A. Visser
Faculteit Wijsbegeerte, Universiteit Utrecht

Contents

Preface

It is commonly accepted that there are about five to six thousand languages. For many pairs of languages $\langle X, Y \rangle$, there is no dictionary $X \rightarrow Y$ or $Y \rightarrow X$, there are only dictionaries for the pairs $X \rightarrow$ English/French/Spanish and dictionaries English/French/Spanish $\rightarrow Y$ (as well as the other way around). There is a clear need for dictionaries translating between languages without the intervention of a small number of Western European languages with a colonial past. Also from a theoretical point of view, such a need can be defended.

The creation of a dictionary of good quality takes a lot of time, and given the fact that 5000-6000 languages yield 25-30 million pairs of languages, it is important to have a database that provides the possibility to translate directly between pairs of languages, even though just a small subset of the millions of pairs will be spoken by sufficiently many speakers to make the enterprise probable. However, sufficiently many will remain to justify a theoretical attempt to think about a database. This thesis highlights some problems that play a role in the creation of such a database, attempts to solve some of them, and tries to show that some other problems cannot be solved.

A well-known problem is that words are often hard to match across languages: different words from different languages do not have the same range of meanings, not all words from one language have an equivalent in the other, etc. In this thesis, a sketch will be given of a database in which most of these problems are solved. Crucial in this set-up is the structure of the interlingua, which provides the possibility to relate non-corresponding meanings in a structural way.

With the set-up proposed in this thesis, which is called SIM*u*LLDA (short for **S**tructured **I**nterlingua **Mu**lti**L**ingual **L**exical **D**atabase **A**pplication) it is possible to generate a descriptive translation for words in the source language that lack a direct translation in the target language. This should ease the work of a lexicographer making a dictionary for a new pair of languages.

In the first chapter, a global explanation is given of why a set-up of a multilingual lexical database that can generate translations between arbitrary pairs of languages is an important asset to have. Furthermore, it

is explained what requirements a lexical database has to fulfil in order to meet this demand. This is done by examining a number of existing theories, and by showing why they are (currently) not capable of generating translations.

The second chapter explains the layout of the alternative lexical database set-up that is proposed in this thesis. As said above the crucial component of the SIM*u*LLDA set-up is the structured interlingua. The structure of the interlingua is provided by a logical formalism called Formal Concept Analysis. Therefore, the chapter gives an introduction of the FCA framework, after wich an explanation is given of how the FCA framework is applied to dictionaries in the SIM*u*LLDA set-up. It is also shown that with this set-up, it is possible to generate definitions for words without a 'translational synonym'. Since FCA is a logical framework, some of the logical properties are discussed, especially those relevant for the functioning of the SIM*u*LLDA system. Additionally, an on-line application called JaLaBA is introduced. This application can be used to generate the Hasse diagrams that represent the structure yielded by FCA automatically, displaying them 3-dimensionally.

In the third chapter, the basic component of the SIM*u*LLDA set-up are investigated in depth: words, languages, interlingual meanings, and definitional attributes. There are two motivations for this thorough investigation. The first is simply that a good theory should explain the intended interpretation of its basic elements, and especially so in the case of an abstract framework like FCA, which does not in itself have any interpretation at all. So if meanings are assigned in the system to *words*, it is important to define precisely what counts as a word. The second motivation is that in the SIM*u*LLDA set-up, meanings are interlingual and defined by means of what might be viewed as 'semantic primitives'. Without restrictions on the interpretation of the primitives (the definitional attributes), the system would make incorrect claims.

In the fourth chapter, the system is empirically tested. As said before, the SIM*u*LLDA system is designed to provide a practical aid for lexicographers. So its practical applicability is of vital importance. The empirical test in chapter four is only a relatively small scale test: to really test the system, it should be applied on a larger scale. But the smaller test in this thesis does show the applicability of the system to an arbitrary semantic domain, and raises a number of important issues.

In the fifth chapter, some extensions to the system are discussed. As discussed in the first four chapters, the system mainly applies to concrete nouns, and then also to the semantic definitions of these concrete nouns. To be a fully workable system, SIM*u*LLDA also needs to model the other components of dictionaries. So, chapter five discusses the treatment of labels, collocations, inflections, derivations, and illustrative examples, also taking into consideration the applicability of the system to other word classes like

abstract nouns, verbs, and adjectives. Finally, some aspects of an application for the system that should be developed in order to use SIM*u*LLDA are discussed.

Background

To understand some of the choices made in this thesis, it is good to know its background setting. The position I held as a PhD student writing this thesis was due to a cooperation between the Utrecht Institute of Linguistic - OTS and the Opleidingsinstituut CKI (the 'department' of Artificial Intelligence). CKI is an interdisciplinary study, in which the traditional disciplines of cognitive science participate: philosophy, linguistics, psychology, and computer science. The original set-up of my research proposal was designed to reflect this multidisciplinary character, and the goal was to give a model of word-meaning that was at the same time psychologically plausible, philosophically feasible, and linguistically applicable. Therefore, there were three people to supervise the project: a linguist (Henk Verkuyl), a philosopher/logician (Albert Visser), and a psychologist (René van Hezewijk).

However, it became quickly apparent that this original goal was hopelessly overambitious: there is no current lexical semantic theory that truly meets any of these three individual requirements, so having a theory that meets all three of them is not feasible for the moment. Psychologically, there are not enough data about the functioning of the brain to really falsify any concrete theory. Linguistically, all lexical semantic theories only have a very resticted domain of application and often fail dramatically when applied elsewhere. And philosophically, it is dubious whether meanings exist at all, or merely emerge in the interaction between the body and its environment.

As a result, the research shifted to taking a given model of word meaning and try to find a useful application for it. This existing model was the only model of word meaning that has stood the test of time: dictionary meanings. So in this thesis, dictionary meanings are taken seriously, and the system tries to really exploit the meaning given in dictionaries rather than improve on the content of their definitions. This is because no matter how much criticism there is on dictionary meanings, no existing lexical semantic theories can escape from this sort of criticism. So extending dictionary meanings with parts of other lexical semantic theories, such as Qualia roles for instance, only multiplies the problems for the theory itself.

In spite of the change of course, the present thesis still has a multidisciplinary character: it discusses issues from a range of fields, including psychology, logic, computer science, philosophy, and last but not least lexicography. The thesis aims at being readable for scientists from all these

different fields. Therefore, as little background knowledge as possible has been assumed, and sections that do depend on a certain field, such as the logical properties of FCA section, are incorporated in such a way that the rest of the text does not depend on it.

This thesis discusses a wide range of topics, all of which are related to the problem of multilingual lexical databases. I hope that it will be a nice introduction to the field and prove interesting to a wide audience. But most of all, I hope it will prove relevant for two groups of people: logicians and lexicographers. For logicians, this thesis will show the force and elegance of Formal Concept Analysis. Also, it gives an illustration of the applicability of the FCA framework to a practical problem. The on-line JaLaBA application described in section 2.5 should also prove interesting. JaLaBA is already generating a good amount of web traffic at the moment.

For lexicographers, it could be even more relevant. As said above, the SIM*u*LLDA system can provide a lexical description for words with a lexical gap. It is useful to have such a description, and as far as I know, there is no other set-up for a lexical database able to generate such a description. In this thesis, I try to show that in order to generate such a description for arbitrary pairs of languages, a multilingual lexical database has to have the following features: it should have an interlingua, this interlingua should be (hierarchically) structured, and the interlingua should contain a representation of differentiae specificae. There might be other ways of getting such a set-up, but it is shown in this thesis that the application of Formal Concept Analysis in SIM*u*LLDA provides an elegant and powerful way of arriving at such a structure.

Chapter 1

Multilingual Lexical Databases

1.1 Multilingual Lexical Databases

There are a lot of different bilingual dictionaries available in the world. Still, that does not mean that there is a bilingual dictionary for every pair of languages. If you consider two 'minor' languages like Malay Indonesian and Hungarian, there is a very slim chance that you will find a dictionary translating between these two languages. This is not surprising: there are several thousand different languages in the world today, so a full coverage would require many millions of bilingual dictionaries.

As a result, the Hungarian who visits Indonesia and wants to understand the language, is left without a dictionary to help him. If he masters the English language, he can find a Malay-English dictionary, in which he can find at least an English translation he might understand. But even if he's very fluent in English, he is still bound to come across English translations he doesn't understand: few non-native English speakers will know what a *rhomboid* or an *ungulate* is, or have an exact grasp of flora and fauna terms like *ivy* or *starling*, other than that the first is a plant, and the second a bird. In these cases, he will need an English-Hungarian dictionary (also readily available), which will give him the translation he needs. But needing two dictionaries makes life unnecessarily complicated.

Nowadays, many dictionaries are available electronically. And if both a Malay-English and an English-Hungarian dictionary were available electronically, one could imagine a computer to first give the English translation of a desired Malay word from the Malay-English dictionary, and then the Hungarian translation of this English translation that it has found in an English-Hungarian dictionary. So what in fact you would like is to automatically link up two bilingual dictionaries in order to get a third one.

But as lexicographers have been aware of for years, you cannot simply link up bilingual dictionaries in this fashion. Of course, Malay-English and English-Hungarian dictionaries will be very useful if you want to create

a Malay-Hungarian one, but they will not simply give you one. In this chapter, I will give a number of reasons why not.

Since creating a bilingual dictionary for every language pair is not a viable option, we need a way to reach the same effect as we would have when we *could* link up bilingual dictionaries: we need to have a system that overcomes all the difficulties preventing the linking of dictionaries. That will be the main purpose of this thesis: to explore an electronic system (a Multilingual Lexical Database, henceforth MLLD) that yields bilingual dictionaries for every pair of languages present within the system, without them having been created pair wise.

Such a system is more difficult than it might seem at first sight. Therefore, this chapter will start out by describing current approaches to multilingual dictionaries and show what fundamental problems they face. This will give us an idea of the basic requirement for the Multilingual Lexical Database that will be proposed in this thesis called SIM*u*LLDA (Structured Interlingua MultiLingual Lexical Database Application).

1.2 Current Approaches and their Shortcomings

This section will describe some current approaches to multilingual lexical databases and their shortcomings, in order to get a clear idea of the requirements an ideal MLLD should fulfil. Since MLLD's are all about words, a good definition of what words are would be necessary in order to give a proper discussion of the precise problems of these current approaches. Such a detailed account will only be given in section 3.1. Until that time, the word 'word' will be used somewhat loosely, making the following discussion less precise than it could be. The only characterisation of words that will be taken into account is that words are not, as de Saussure would have them, seen as pairs of form and meaning, but words are seen as more abstract entities, where the same word can have various meanings. Also, no strict distinction will be made between meanings and senses.

1.2.1 Parallel Wordlists

A word of a certain language, say Dutch, can not only be translated into a single other language, say English, but into many other ones. Therefore, there is a certain logic in creating a single dictionary which contains all these translations, instead of having a new book for every pair of languages. And indeed, at least in Europe, there are several of such Multilingual Dictionaries, providing translations between the major European languages. For each language in such a dictionary, there is a separate part, in which the words are alphabetically ordered by the words of that language. An example is the 'Euro Dictionary' by Goursau, for Dutch, English,

French, German, Spanish and Italian. It consists of six parts, a sample of two of which is given in table 1.1.

NL	Fr	En	Ge	Sp	It
bang (zijn)	peur (avoir)	afraid (to be)	Angst (haben)	miedo (tener)	paura (avere)
bang maken	effrayer	frighten	erschrecken	asustar	spaventare
bank	banc	bench	Bank (f)	banco	panca; panchina
bank	banque	bank	Bank (f)	banco, banca	banca
bank-	bancaire	bank, banking	Bank-	bancario, a	bancario, a
bankbediende	employé, e	clerk	Angestellte (f,m)	oficinista	impiegato, a
bankbiljet	billet (banque)	bank-note	Geldschein (m)	billete (de banco)	banconota
banketbakkerij	pâtisserie	cake-shop	Konditorei (f)	pasteleria	pasticceria
En	**Fr**	**Ge**	**Sp**	**It**	**NL**
banister, handrail	rampe	Geländer (n)	barandilla	rampa	trapleuning
bank	banque	Bank (f)	banco, banca	banca	bank
bank, banking	bancaire	Bank-	bancario, a	bancario, a	bank-
bank	rive	Ufer (n)	orilla, margen	riva	oever
bank	berge	Ufer (n)	ribera, orilla	sponda	oever
banker	banquier	Bankier (m)	banquero	banchiere	bankier
bank-note	billet (banque)	Geldschein (m)	billete (de banco)	banconota	bankbiljet
bankrupt (go)	faillite (faire)	bankrott sein	quiebra (hacer)	fallire	failliet (gaan)

Table 1.1: a Sample of the Goursau Euro Dictionary

The idea behind a dictionary like this is, that if you put a Dutch word next to its translations of various other languages, you can translate between all these languages within one volume. Now in its printed version, this set-up has the afore mentioned drawback that only one of the languages can be put in an alphabetic order, so that a separate part has to be created for each language present in the multilingual dictionary. But firstly this is still much less than the 30 different parts[1] one would have needed otherwise, and secondly, this problem simply disappears if we consider an electronic version of the dictionary. For in an electronic version, sorting is something that can be done on the fly, so that a single file would suffice, which could be sorted on any of the languages. Furthermore, electronic dictionaries are usually not browsed in the traditional fashion, but are more often consulted using a query, so that no sorting is needed anyhow.

However, multilingual wordlists face a more serious problem: the implicit assumption in such a set-up is, that a word in one language corresponds fully to a word in another language: that a word has a true synonymous word in that other language, and in all other languages. So, if you

[1]For each of the 6 languages you would need one to translate to the other 5.

pair up all these translational synonyms, you can translate words from one language to another.

But words are not the kind of things that equate between languages, as can easily be illustrated using table 1.1: the second half of the table contains various entries for the English word *bank*, such as one in which the French word banc is given as the translation, and one in which the French word rive is given. If the relation between the words in the rows in table 1.1 would be one of identity, that would imply that the French words rive and banc would have to be identical, which they are clearly not. So the conclusion has to be that either the first English word *bank* is not identical to the second one (which is an unfortunate choice, further discussed in chapter 3), or that the words *bank* and *rive* are *not* truly identical to each other.

To make the same point again: take the entry for *bank* in the van Dale English-Dutch dictionary (table 1.2) and the Dutch-French entries of the Dutch translational synonyms mentioned therein (table 1.3)[2]. Equating words across languages would mean that, since the French word *banc* is assumed to be identical to the Dutch word *bank*, and the Dutch *bank* is in turn the same as the English *bank*, the French word *banc* should translate into the English word *bank*, which it does not. And neither do the French words *ourlet*, *rangée* or *réserviste*: *banc* is *bench* in English ; the word *ourlet* translates into *hem*; *rangée* means *file* or *line*; and a *réserviste* is a *reservist*.

bank[1] [bæŋk] ⟨telb.zn.⟩ **0.1** ⟨ben. voor⟩ ***bank*** ⇒ *mistbank; wolkenbank; sneeuwbank; zandbank; ophoging; aardwal* **0.2** ***oever*** ⇒ *glooing; talud* **0.3** ***bank*** ⟨ook als gebouw⟩ ⇒ *geldbedrijf* **0.4** ***reserve*** ⇒ *voorraad; spaarpot* ⟨ook lett.⟩, *bank* ⟨in kansspelen bv.⟩ **0.5** ⟨ben. voor⟩ ***rij*** ⇒ *serie; batterij* **0.6** ⟨lucht⟩ ***slagzij*** ⇒ *dwarshelling* **0.7** ***boord*** ⟨v. biljarttafel⟩ **0.8** ⟨scheep.⟩ ***doft*** ⇒ *roeiersbank* **0.9** ⟨mijnw.⟩ ***schachtmond*** ⇒ *putrand*

Table 1.2: van Dale English-Dutch definition of *bank*

bank **0.1** [meubel, ook in kerk, school, buiten] ***banc*** ⟨m.⟩; ⟨in auto, trein⟩; ***banquette*** ⟨v.⟩; ⟨in huis ook⟩ ***canapé*** ⟨m.⟩ **0.2** [instelling; gebouw] ***banque*** ⟨v.⟩ *0.3* [casino] ***casino*** ⟨m.⟩ *0.4* [werkbank] ***établi*** ⟨m.⟩ *0.5* [zandbank] ***banc (de sable)*** *0.6* [inzet] ***banque***

boord **0.1** [aan kledingstuk] ***bord*** ⟨v.⟩ **0.2** [zoom] ***ourlet*** ⟨v.⟩ **0.3** [kraag] ***col*** ⟨m.⟩ ⇒ ⟨losse kraag⟩ *faux col* **0.4** [scheepswand; schip, luchtvaartuig] ***bord*** ⟨v.⟩

oever **0.1** ***bord*** ⟨v.⟩ ⟨niet zelfstandig te gebruiken⟩ ⇒ ⟨van meer, grote rivier, enz.⟩ *rive* ⟨v.⟩

reserve **0.1** [alg.] ***réserve*** ⟨v.⟩ **0.2** [sp.] ***ramplaç|ant*** ⟨m.⟩ ***-ante*** ⟨v.⟩ **0.3** [mil..] ***réserviste*** ⟨v.⟩

rij **0.1** [opeenvolging in rechte lijn] ***rangée*** ⟨v.⟩; ⟨achter elk.⟩ ***file*** ⟨v.⟩ ⇒ ⟨van personen ook⟩ ***queue*** ⟨v.⟩ **0.2** [volgorde] ***suite*** ⟨v.⟩ **0.3** [reeks] ***série*** ⟨v.⟩

Table 1.3: some Dutch-French definitions

[2]The definitions of *slagzij* (banking plane), *doft* (rowing bank) and *schachtmond* (bank of a shaft) are left out, since the van Dale Dutch-French has no entries for them.

It is obvious what goes wrong here: though the English word *bank* translates into the Dutch word *bank*, it only relates to a certain *meaning*[3] of it; it does not really take on any other meaning the Dutch *bank* might have, but just some (or even only one) of them. In this case, both can stand for a financial institution, and both can also name a specific sort of pile or ridge of stuff, like in a sandbank, or a bank of clouds. But whereas the Dutch word *bank* can also name a bench, the English word can not, and where the English word *bank* does also stand for the side of a river, the Dutch word doesn't[4].

So it is too simple an idea to try to link words across languages on the bases of equivalence; words of different languages may have similar meanings, but they are not simply identical (although they can of course share *all* their meanings). The result of this is, that any method that tries to link languages at the level of words (as wordlists do) will fail as soon as it has to deal with ambiguous words that do not coincide on all their meanings across languages. Thus, a proper multilingual lexical database has to account for the fact that when two words (of different languages) have a similar or identical meaning, that does not make them (translationally) identical words.

1.2.2 Hub-and-Spoke Model

In order to overcome the problem with ambiguous words, languages have to be linked not at the level of words (or word-forms), but at the level of meanings. And in order to do that, a strict distinction between words and meanings has to be made. This is the approach taken by, for instance, the Linkable Resource Lexicons (LRL's):

> If the objects to be linked, as in our case, are 'words', then a clear distinction should be made between words as form units (the form of the word) and words as meaning units (the meaning of the word). Otherwise it should be the case that, having linked the English form knight both with the French form chevalier and with the German form Ritter, chevalier and Ritter also would be linked to knight in its meaning as a chess piece which would lead to an incorrect link both for French (where one should find cavalier for the chess piece), and for German (where the correct link is with Springer or Pferd).
> (Beeken *et al.*, 1998 [22])

[3]Or sense, we will not make a distinction between these at the moment.

[4]It should be said that by the fact that there are various different entries for the English word *bank* in the Goursau dictionary, this problem has in some way been accounted for: the financial institution *bank* and the riverside *bank* are put on different rows. But the problem is, that there is no indication as to which is which: there is no way of telling the different meanings of *bank* apart.

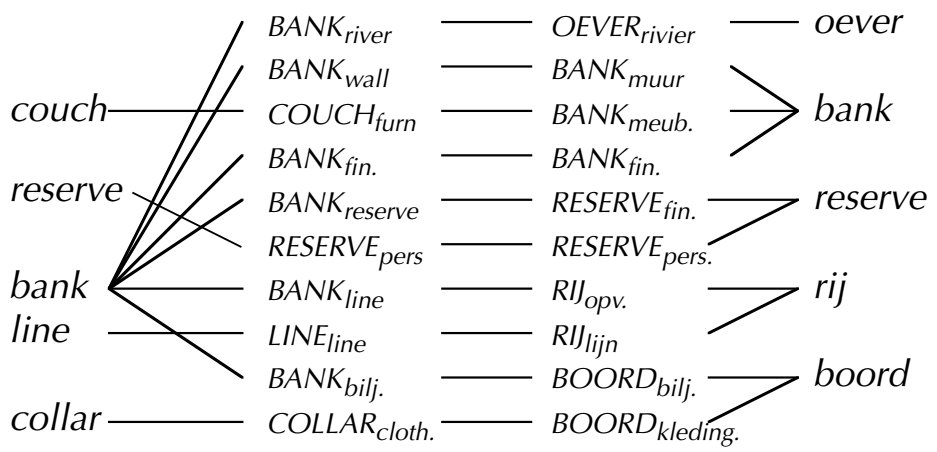


Figure 1.1: OMBI structure for *bank*

A system that uses LRL's is OMBI[5]: a tool for creating bilingual dictionaries, developed by the software-house SERC under the auspices of the Dutch-Belgian CLVV Committee (Commissie voor Lexicografische Vertaal-Voorzieningen[6]). OMBI aims at simplifying the creation of new dictionaries by pursuing a number of properties: language pairs can be reversed; different dictionaries can be derived from one database; additional languages can be easily added (thus creating multilingual resources); and all this is attempted in a generic and non-language-specific way (Beeken *et al.*, 1998 [62]).

The way the OMBI system is set up (using the pair English-Dutch as an example), is the following. There are four individual 'layers' within the system (presented vertically in figure 1.1): two lexical layers – one for Dutch words (Form Units or FU's), and one for English words – and two *conceptual* layers – one containing the meanings of the Dutch words (Lexical Units or LU's), and the other the ones for the English words.

The different languages are not linked at the level of the FU's, but at the level of their meaning, hence with their LU's. To give an example: the English word *bank* is related, amongst others, to an English 'river-meaning', which is equated to an equivalent notion in the Dutch conceptual layer, which in turn is related to the Dutch word *oever*. The Dutch and the English words can have different meanings related to them, and different FU's can also relate to the same LU. Hence, the system can get rather tangled, as illustrated in figure 1.1.

[5]Omkeerbaar Bilinguaal = Reversible Bilingual
[6]Commision for Lexicographic Translation Tools

Without further refinements, this set-up is still not precise enough to correctly connect languages. Although it does not assume identity of words across languages, it *does* assume that the meaning of a word corresponds to a single meaning of another word in another language. And this too leads to problems in some cases: the Russian word *голубой* (goluboj) means *blue* in English. But it specifically means *light blue*; for dark blue there is another word, namely *синий* (siny). English doesn't make this distinction (not lexically that is) and has only the word *blue*. And where English and Dutch have different words for fingers and toes, Spanish and Italian only have a more general term *dedo* (or *dito* respectively) covering both fingers and toes.

What goes wrong with the treatment of such terms in OMBI is that there are no appropriate LU's to identify between the languages: neither the more general, nor the more specific LU's can be related to an equivalent LU in the other language. A possible solution to this problem is to introduce LU's for light- and dark blue in the English conceptual layer, and have both of these concepts relate to the FU blue. That is the solution adopted in, for instance, the Conceptual Structures architecture proposed by Sowa (1993). This effectively introduces a polysemy into English on the basis of the following principle:

> Si A de L_1 est équivalent à α de L_2 et si A de L_1 est équivalent à β de L_2 alors que α de L_2 n'est pas synonyme de β de L_2, c'est que probablement A de L_1 possède deux sens qui devraient être différenciés par deux noeuds du réseau[7]. (van Campenhoudt, 1994 [64])

However, this is a conceptually unattractive move: the fact that Russian has a different set of lexicalised meanings should not lead to the introduction of new meanings in the English language[8]. The alternative is to reduce the power of translational equivalence: *"as a general rule, meanings of two LUs that are translation equivalents are not identical, which means that translation equivalence is basically approximate"* (Mel'čuk & Wanner, 2001 [25]). Mel'cuk & Wanner call cases such as *blue* above cases of *multiple lexical correspondence*, of which they claim that they *"can be relegated neither to the source not to the target language alone. They have to be faced 'between' languages"*. (Mel'čuk & Wanner, 2001 [26]).

The solution that is adopted in the OMBI system is to have various possible relations between LU's of different languages. Beside the strong relation of equivalence, LU's of different languages can also be linked by weaker relations such: *hyperonym*, *hyponym*, *near-equivalent*, and even the very weak

[7] If A of L_1 is equivalent to α of L_2 and if A of L_1 is equivalent to β of L_2, whereas α of L_2 is not a synonym of β of L_2, then A of L_1 probably has two meanings that should be differentiated by means of two distinct entries within the dictionary. [translation by van Campenhoudt]

[8] Although van Campenhoudt (2001) argues that this introduction is motivated, because of the existence of *hyponomase*: the mechanism by which we use a hyperonym to anaphorically refer to a hyponym.

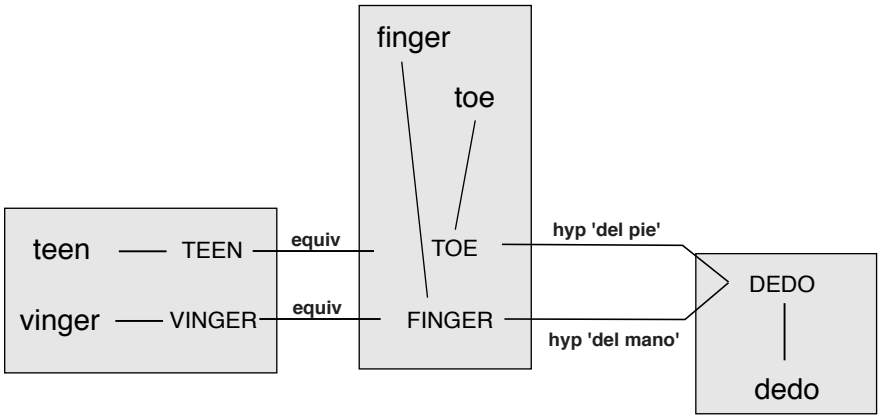


Figure 1.2: Example of the Hub-and-Spoke model

relation *related*. Thus, the Spanish LU dedo has a hyponym-link to both the English finger and the English word toe. Furthermore, both hyponym-links have semantic restriction: the former as ⟨del mano⟩ and the latter as ⟨del pie⟩.

The LRL approach is not a multilingual set-up: it simply defines a link between two languages. So though it contains many ideas to make the creation of dictionaries more efficient (i.e. reversibility and reusability), OMBI still links languages pairwise. There is, however, a multilingual extension to the LRL approach, called the Hub-and-Spoke model, developed by the IMS[9] (Stuttgart) and the CLVV, under supervision of W. Martin. In the Hub-and-Spoke model, one of the languages is selected as the central language. This central language can then be used as a *'hub'* to connect languages that are not directly linked[10], by exploiting the links they both have to the hub. So if Dutch were the hub, then the situation depicted in figure 1.2 would arise: the Dutch LU vinger would be linked as an equivalent to the English LU finger, while teen is equated with toe. Now because of the link of the English LU's to the Spanish LU dedo, we could deduce that a vinger is a "dedo del mano", and a teen is a "dedo del pie".

In short, the Hub-and-Spoke model uses the LU's of a particular language (the hub) as the basis for linking languages, where the structural information regarding hyperonymy is represented in the links between the languages.

[9]Institut für Machinelle Sprachverarbeitung=Institute for Computer Language Processing

[10]The possibility of having several central hubs is explicitly mentioned, but it does not fundamentally change the following points.

As observed by Vossen et al., choosing one of the languages as a central linking tool has a considerable disadvantage: *"linking ... through one of the languages ... forces an excessive dependency on the lexical and conceptual structure of one of the languages involved."* (Vossen *et al.*, 1997 [1]). To translate this to the *finger/toe* example : if Spanish were taken as the hub, a problematic situation would arise. Since Spanish does not have a LU for 'toe', the two equivalent LU's TOE and TEEN would have to be linked via the more general Spanish LU DEDO. This would not relate these two LU's as *equivalent*, as should be, but as hyponyms of hyperonyms.

This problem is not beyond repair in the Hub-and-Spoke model: the Dutch and English words will not be simply linked as hyponyms of *dedo*, but as hyponyms with a distinctive feature: either 'del mano' or 'del pie'. And those hyponyms with identical distinctive features (i.e. *teen* and *toe*) can than be reconstructed as being equivalent. But there are two problems with this solution: firstly, there is the pragmatic problem that since the distinctive features are themselves language-dependent text-strings, their identity will be hard to establish. But more importantly, it is a solution to a problem that should not have existed in the first place, by forcibly assigning a heavy function to those distinctive features.

A much more elegant solution would be to not have one of the languages to function as a hub, but to have an independent interlingual structure connecting the various languages: an interlingua aligning the individual languages at the level of their meaning. One of the major projects using such a setup is the *EuroWordNet* project, which will be discussed in the next section.

1.2.3 WordNet and EuroWordNet

EuroWordNet is a large project, mostly financed by the European Union, which aims at a multilingual lexical database connecting various independently created language 'nets' through an unstructured set of Interlingual Lexical Items (ILI's). It is a multilingual version of the Princeton WordNet project. Before turning to the multilingual extension, I will first describe the original, monolingual WordNet project.

WordNet

In the seventies, the main approach to lexical semantics was based on the theory of Componential Semantics, defended for instance by Katz & Fodor (1963). Componential Semantics held the claim that *"the meaning of a sentence should be decomposable into the meaning of its constituents, and the meaning of a word should be similarly decomposable into certain semantic primitives, or conceptual atoms."* (Miller, 1998 [xvi]). These semantic primitives can then

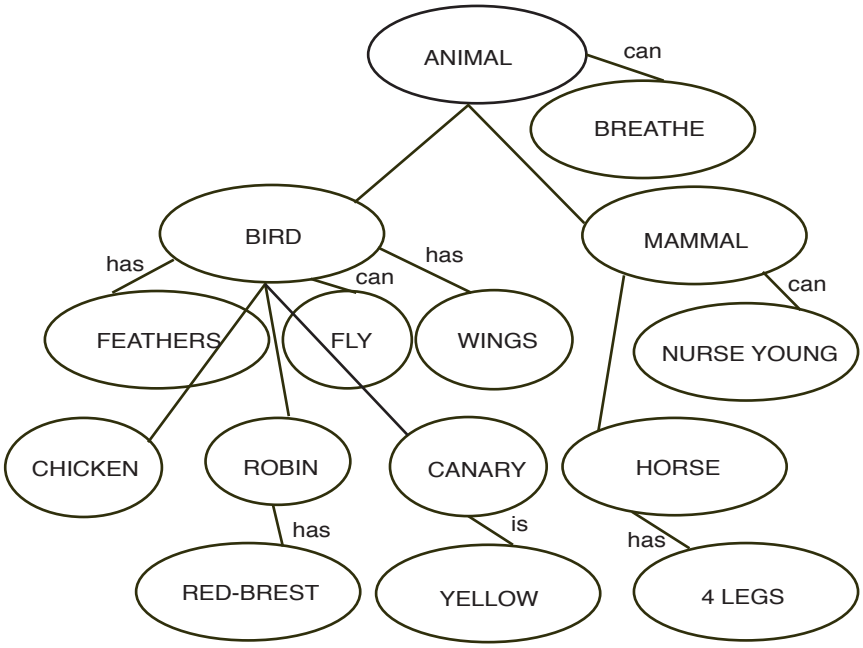


Figure 1.3: A Semantic Network (Collins & Quillian, 1969)

be properly founded, and serve as a solid basis on which the meaning of all the complex terms can be founded. (more on this in section 3.4.1).

The central problem of componential semantics is that it is by no means clear that there are such semantics primitives, and if there are, which these should be. In the light of this, George Miller proposed WordNet as an alternative to these decompositional approaches to the lexicon, based upon the psychological Semantic Network Theory.

Semantic Network Theory, originally developed by Quillian (1968) but better known in its later versions first by Collins & Quillian (1969) and Collins & Loftus (1975)[11], is a theory in which the meaning of a word is not dependent on a small set of primitives, or some external foundation, but only on other words: the meaning of every word is completely determined by its relation to other words. Typically, such relation are **is_a**, **has**, and **part_of**. An example of a part of such a semantic network is given in figure 1.3.

In a naive interpretation, the WordNet approach would immediately run into troubles when ambiguous words are considered. The basic problem is similar to that of the parallel word lists: the relations between words are

[11] A nice overview of semantic network theories can be found in Johnson-Laird et al. (1984).

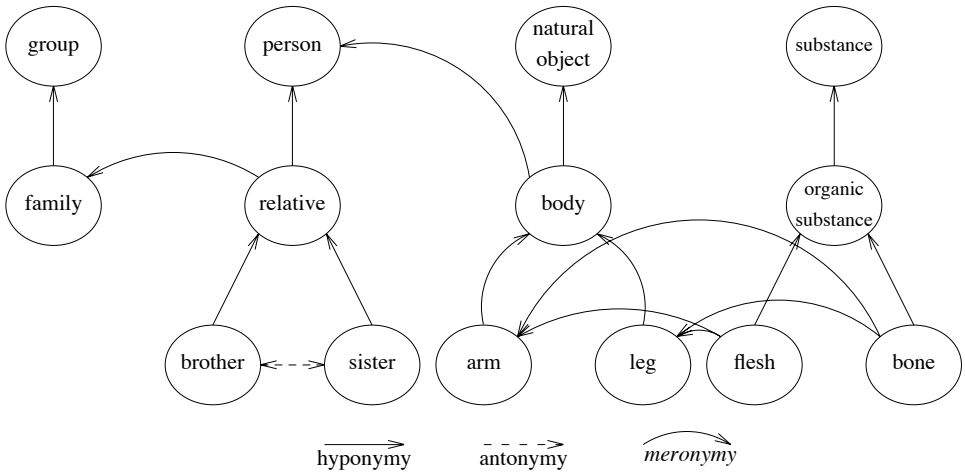


Figure 1.4: Example of a Wordnet Structure (Miller, 1990)

not at the level of the word, but rather at the meaning level; it is not the *word horse* that has an '**is_a**' relation to *animal*, but only one of the senses of that word. Hence, WordNet does not use words as elements in its semantic networks, but rather *synsets*: sets of synonymous word-meanings. And since the same word can partake in various synsets, this effectively solves the ambiguity problem.

Individual synsets are connected, as in Semantic Networks, by means of lexical relations. For nouns, the set of lexical relations consists of hyponymy, hyperonymy, meronymy, and antonymy. These relations are assumed to sufficiently characterize the place of the synset in the lexical field and hence determine its meaning. An example of a part of a wordnet is displayed in figure 1.4.

EuroWordNet

The idea behind the EuroWordNet project is to create an individual wordnet for each of the (major) European languages. Synsets within the individual wordnets are linked by the same lexical relations as used in WordNet, whereas an additional relation **eq_synonym** assures that synsets can also be linked cross-linguistically, hence resulting in a highly flexible multilingual lexical framework.

Because of the redundancy of relations that would be needed in case languages would be linked up pairwise, the translation **eq-synonym** is not defined between the synsets of the various languages directly, but rather between the synsets of a language, and a set of interlingual items (ILI's).

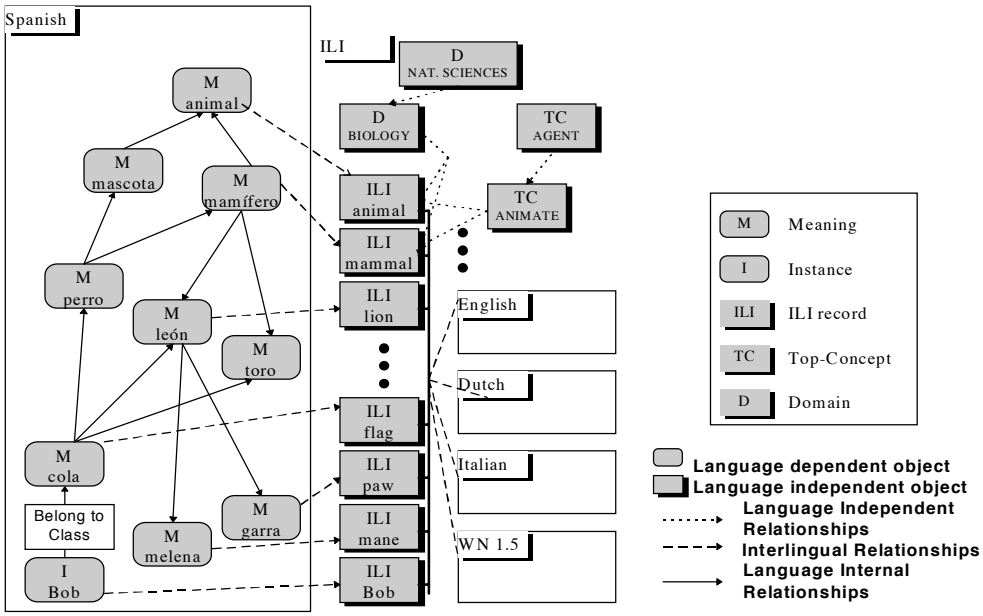


Figure 1.5: The Setup of EuroWordNet (Vossen *et al*, 1997)

It would be possible for these ILI's to be highly structured by themselves. However, for pragmatic reasons, this was not done within the EuroWordNet project: "*A language-independent conceptual system or structure may be represented in an efficient and accurate way but the challenge and difficulty is to achieve such a meta-lexicon, capable of supplying a satisfactory conceptual backbone to all the languages.*" (Vossen *et al.*, 1997 [1]). So the interlingua in EuroWordNet consists simply of a list of ILI's, with no internal structure by itself. The main function of these ILI's is to function as a 'hub' in the sense of the hub-and-spoke model:

> Each synset in the monolingual wordnets will have at least one equivalence relation with a record in this ILI ... Language-specific synsets linked to the same ILI-record should thus be equivalent across languages. The ILI starts off as an unstructured list of WordNet 1.5 synsets, and will grow when new concepts will be added which are not present in WordNet 1.5. (Vossen *et al.*, 1997 [2])

The overall structure of EuroWordNet is illustrated in figure 1.5. Beside the wordnets and the ILI's, there are two additional types of entities: top-concepts and domains, both providing a rough classification on ILI's. For present purposes, however, these can be ignored.

If the synsets in the language modules would be linked to the ILI's only by use of synonymy-relations, the problem of concepts that do not fully

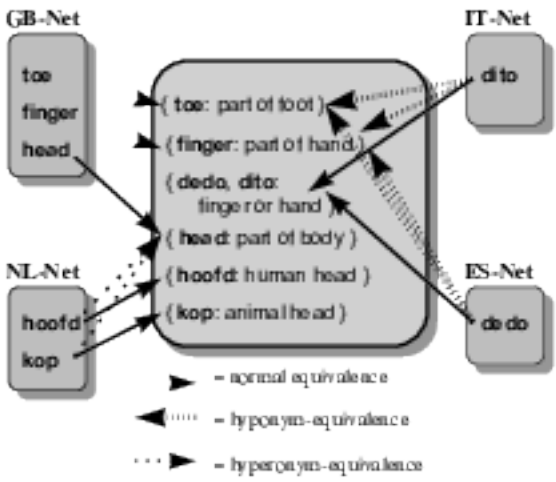


Figure 1.6: Fingers and Toes in EuroWordNet (Vossen, 1997)

correspond across languages, like the *finger/dedo* example, would remain (unless we accept the non-translatability of non-corresponding terms): although for both *finger* and *dedo* there would ILI's with which they have an equivalence relation, there would be no relation between these ILI's, and hence no relation between these words.

As a solution to this, EuroWordNet proposes a number of additional relations beside the **eq_synonym** relation, most notably: **eq_near_synonym**, **eq_has_hyperonym**, and **eq_has_hyponym**[12]. Using these relations, the problem with the single word for *finger* and *toe* in Spanish can be resolved in the following way: first, three independent ILI's are defined: FINGER, TOE and DEDO. All the English and Spanish words are related to all three of them: *dedo* has a **eq_synonym** relation to DEDO, and also a **eq_has_hyponym** relation to both FINGER and TOE, whereas *finger* and *toe* are related with **eq_synonym** relations to FINGER and TOE respectively, and both with a **eq_has_hyperonym** to DEDO.

All this results in a very elegant set-up, illustrated in figure 1.6. What makes it especially elegant is that the special relation **eq_has_hyperonym** only has to be exploited to link languages which do in fact have non-corresponding terms: the nicely translational synonymous terms such as *toe* and *teen*, and *dedo* and *dito* can simply be directly linked with **eq_synonym** relations, with the more complex relations no longer playing a role.

But despite its elegance, this set-up faces two fundamental difficulties, the first relating to the lack of structure on the ILI's and the multiplicity of

[12]The set of relations is in principle open, and includes functional relations like **eq_has_holonym** (indicating meronymy), and **eq_is_subevent_of**.

structures elsewhere, and the second relating to the lack of semantic restrictions.

There are various places in which the meanings in EuroWordNet are ordered: every language in the EuroWordNet system has its own wordnet, with its own lexical relations between synsets, the **is_a** relation of which instantiates a hierarchical order. The synsets themselves are in turn possibly hierarchically related to the ILI's, with the relations **eq_has_hyperonym** and **eq_has_hyponym**. The ILI's by themselves are not ordered, but are related to Top Concepts. Top Concepts define a general ordering on ILI's, and have by themselves a hierarchical ordering. All these different orderings are supposed to order more general terms over more specific ones.

Since all these orderings order on the same aspect (generality), they should all run in parallel: to have it otherwise would be the same as having two numeral systems, say Arabic and Roman, both ordered, but then say that iv $<$ vi does not imply that $4 < 6$. Hence, because the English word *dog* is a hyponym of *animal*, the French translation of the first (*chien*) has to be a hyponym of the French translation of the second (*animal*).

But if all these orderings are necessarily parallel, there is in fact only one ordering, although this ordering is present at distributed places throughout the system. Thus, the ILI's *are* implicitly ordered, even though they are represented as having no structure: the fact that the English synset for *dog* is a hyponym of the synset for *animal*, means that the ILI DOG is a 'hyponym' of the ILI ANIMAL. So not having order on ILI's has little theoretical merit, and is only less clear. There is even a risk in such a set-up: there is nothing in the system to prevent mismatches between the various orderings, thus allowing hidden incompatibilities between the various parts of the system. Therefore, in the present thesis, the aim is to only have *one* ordering on synsets/words/concepts, which is imposed upon the items in the interlingua.

To turn to the second problem: the hub-and-spoke model labels hyperonymy and hyponymy relations with semantic restrictions. This enables it to say that although the German *Beugung* and *Konjugation* both translate into *inflection* in English, there is a difference between them: *Beugung* applies only to nouns, whereas *Konjugation* applies only to verbs. Since EuroWordNet does not have such decorations on the **eq_has_hyperonym** relations, it cannot keep these two words apart like that; and it only indicates that *finger* and *toe* both are more specific than *dedo*, and that they are not synonymous, but not what is the nature of their difference is.

This problem not only arises in a multilingual setting, but already causes problems for the monolingual WordNet system. Especially higher up in the ordering, there are many synsets in WordNet that have a lot of hyponyms. For instance: *runner*, *smoker*, *sleeper*, and many others are all direct hyponyms of *person*. And without further specification of what is special about a smoker (that he smokes), the system will contain no more informa-

tion than that these are all different terms for "persons". This would not be a problem if the system would have differentiating features on hyponymy links.

This leads to a number of basic requirements that a multilingual lexical database should fulfil: its should keep words and meanings well apart, it should link languages at the level of their meaning, it should link language via a non-language-dependent interlingua, the items in this interlingua should be ordered, and hyponymy relations should be decorated with differentiating features. These will be the benchmarks that the system proposed in the next two chapters will have to face.

1.2.4 Acquilex et al.

The three projects discussed thus far are all what Ooi (1998) calls *human-consumption* approaches: approaches concerning lexicographic information for a human reader. Human consumption is also what the proposed system (SIM*u*LLDA) aims at. But a larger number of computational approaches to lexicography are oriented towards machine use, mostly with a well-defined goal: the problem of machine translation. On the one hand, these projects and their findings are highly relevant for the purpose of this thesis. On the other hand, however, there are fundamental differences between human-consumption and machine-use databases.

Using the Acquilex project as an example, I will try to show these differences in this section, and make clear why this makes the findings of such approaches not directly applicable for human consumption approaches.

Lexicons built for machine translation do not use entries that resemble dictionary definitions: given their purpose, such lexicons try to give a more formalised representation of word meanings, containing information relevant for translation problems. For instance, the Acquilex project tries to represent "*some aspects of lexical semantics ... formally, within a unification-based framework, in a way which integrates with syntax and compositional semantics* ", and to apply "*the results of this ... in the construction of large-scale semantic knowledge bases*" (Copestake, 1992 [1]). This semantic knowledge base that is constructed by Acquilex is called the Lexical Knowledge Base (LKB).

The philosophy behind the LKB is best shown using an example, such as the entry for *chocolate* in figure 1.7. It presents a syntactico-semantic element, with some overlap with a traditional dictionary entry: it says that the orthography of this particular meaning is `chocolate`, and that its lexical category is that of a noun. However, it also contains both more and less information that the lexical entry in a dictionary.

More than a dictionary entry, the LKB contains information that linguistic theory deems necessary for the interpretation and usage of the word

$$\begin{bmatrix} \textbf{lex-noun-sign} \\ \text{ORTH} = \textbf{chocolate} \\ \text{CAT} = \boxed{\textbf{noun-cat}} \\ \text{SEM} = \begin{bmatrix} \textbf{unary-formula-entity-arg1} \\ \text{PRED} = \textbf{chocolate\_L\_1\_4} \\ \text{ARG1} = \textbf{entity} \end{bmatrix} \\ \text{SENSE-ID} = \boxed{\textbf{sense-id}} \\ \text{QUALIA} = \begin{bmatrix} \textbf{c\_art\_subst} \\ \text{PURPOSE} = \begin{bmatrix} \textbf{formula} \\ \text{PRED} = \textbf{drink\_L\_1\_1} \end{bmatrix} \\ \text{PHYSICAL\_STATE} = \textbf{liquid\_a} \\ \text{CONSTITUENCE} = \boxed{\textbf{constituency}} \\ \text{FORM} = \begin{bmatrix} \textbf{physform} \\ \text{SHAPE} = \textbf{non-individuated} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 1.7: The LKB entry for chocolate (Copestake, 1992)

in a sentence. The most striking example in figure 1.7 is the presence of a qualia structure, adopted from the Generative Lexicon theory of Pustejovsky (Pustejovsky, 1995a). Part of the argumentation for the necessity of the qualia structure for the correct interpretation of words is the following: the verb *to start* takes a verb as its argument, so one can start walking or start drinking. But the phrase *to start a stone* is ungrammatical, since it contains an internal argument of an incorrect type. However, it is possible to say *start a book*, or *start a chocolate*[13]. The reason for this, according to Pustejovsky, is that both *book* and *chocolate* have a PURPOSE role in their qualia structure (Pustejovsky calls it a *Telic Role*), consisting of a verb. This verb within the telic role can be dragged out to coerce the noun into a verb. So to start a book gets the (default) interpretation of to start *reading* a book, while 'to start a chocolate' reads as 'to start drinking a chocolate'. Stones, on the other hand, have no such special purpose, and hence cannot get coerced in this fashion.

But there is also information from dictionaries that is absent in Acquilex: what is lacking from this LKB entry is information distinguishing chocolate from other drinkable liquids. There is nothing indicating that whereas apple juice is made from apples, chocolate is made by mixing chocolate with water or milk (as Webster defines it). There is the information that the word *chocolate* can be semantically represented by the

[13] Arguably, the word *start* in *to start a motor* has a different meaning, and in that meaning takes an object as its argument.

predicate **chocolate_L_1_4**, but that results in little more than claiming that chocolate *is* different from other drinks, but not in what respect.

This lack of differentiae specificae is not a sloppy failure of the theory, but a design feature of the Acquilex set-up: like other machine use approaches, Acquilex is directed toward the (computational) production and perception of sentences, and as such is concerned with semantics only in as far as it can be formally expressed, and it has consequences for the usage of the word. And the fact that chocolate, other than tea or coffee, is a solution containing cocoa has no such impact:

> A speaker does not need the information that a rabbit has long ears in order to use the word appropriately. In fact, the only information given in the dictionary definition which clearly fits this criterion is that a rabbit is an animal. (Copestake, 1992 [2])

This claim that much of the content of a dictionary definition is irrelevant, is the crucial reason why machine use approaches are not directly relevant for the current project: SIM*u*LLDA is an attempt to better organise the lexicographic information found in dictionaries. Therefore, the global set-up of an approach that, like Acquilex, rejects lexicographic information as irrelevant, is only indirectly relevant. Many of the problems and solutions described in the Acquilex project are relevant (and will be discussed throughout this thesis), and the very existence of such theories raises a very fundamental question: is traditional lexicographic information relevant? This question will be discussed at length in paragraph 3.3 and 3.4.4. But it is not to be expected that the overall design of a machine use project could be directly used for a human consumption approach like SIM*u*LLDA.

Thus far only the monolingual set-up of machine use projects has been discussed. But many of these projects also have a multilingual component. Naturally, the concerns of reusability and lexical gaps stated previously in this chapter also apply for machine use projects. For instance, the Acquilex team states the following:

> Recently, there has been considerable interest in encoding multilingual transfer rules in unification-based formalisms, aiming for declaritiveness and bidirectionality, but allowing sufficient expressiveness to deal with lexical 'gaps'; specialisation, and so forth ... Our concept of translation equivalence maintains these advantages, but abstracts away from MT transfer rules. (Copestake *et al.*, 1992 [2])

Within most of the computer use projects, the multilingual linking of languages is done by a separate mechanism. For instance, Acquilex has a system of **tlinks** for this purpose. These work as follows: in easy cases, a **tlink** simply links up translationally synonymous feature structures (LKBs) between languages. In case of a lexical gap, the system is slightly more complex as indicated in figure 1.8: the feature structure in the source language

(SFS1) and the feature structure in the target language (TFS1) cannot be directly linked, since they are not equivalent. Therefore, they are mapped onto underlying feature structures SFS0 and TFS0, that are equivalent (by design).

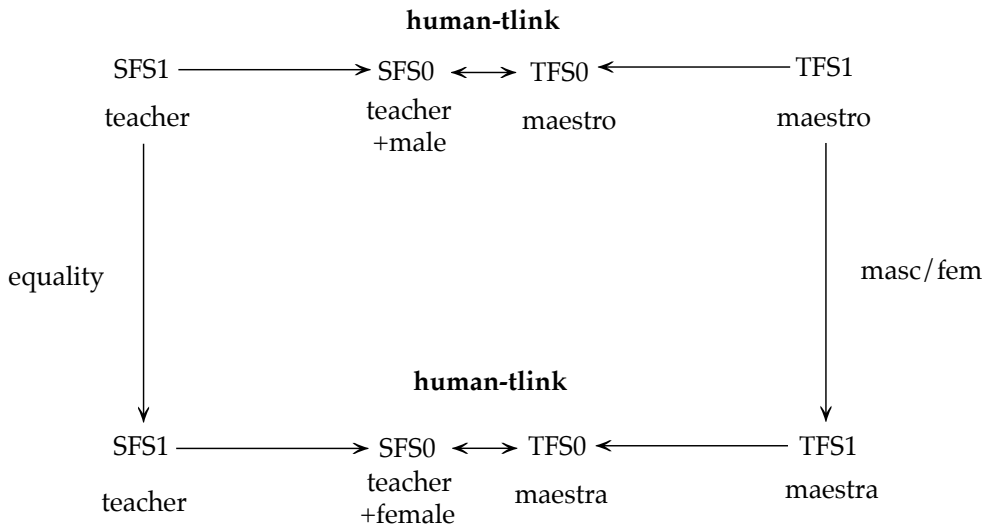


Figure 1.8: The **tlink** for teacher/meastro (Copestake, 1992)

This approach is very close to the mechanism of the Linkable Resource Lexicons: Form Units (SFS1 in this case) are linked via underlying Lexical Units (SFS0). In the case of an absent translational synonym, such as the lack of a word for a specifically male teacher in English, the lexical gap is filled by adding the required notion at the level of the meanings. The only minor difference is that in the case of the Hub-and-Spoke set-up, the same FU is connected to various LU's, whereas in the Acquilex setup a copy of the original SFS1 is made.

Since the Acquilex approach is so similar to the Hub-and-Spoke approach, the same problems also apply: the pairwise linking of languages is not easily extensible to a multilingual setting, and adding meanings to a language that do not belong to it (such as adding male-teacher to English) is not a very elegant solution.

So the reasons for not treating computer use approaches separately in the discussion of the basic requirements of a multilingual lexical database are twofold: on the one hand, as a monolingual lexical representation language, the goals of computers use approaches differ too greatly from that of a multilingual lexical database based upon traditional lexicographic information. And on the other hand, as a multilingual linking tool, they are too

similar to human consumption approaches to provide new insights. This does not mean that the findings of computer use approaches are irrelevant for the present purpose, but the problems encountered by for instance Acquilex will be discussed at various places throughout this thesis.

1.2.5 Corpus Based Approaches

There is another influential brand of multilingual lexical research that should be mentioned here: (multilingual) computer corpus lexicography (henceforth MCCL). Behind corpus based methods lies a very simple idea. As previous failure of various attempts have illustrated, it is impossible to define words by postulating their meaning. The meaning of a word is not something that can be externally decided, but is a property of the word itself, that is brought forth by the way the word is used. So the conclusion should be that "*the only way to 'define' the meaning of* charge *[for instance] is to describe (usually by illustrative phrases or sentences) the distribution of the word.*" (Nida, 1958 [282]).

This claim raises a lot of interesting questions, such as: is it true that the meaning of a word can be given merely by looking at its distribution in a corpus? Is there nothing more to meaning than use? Are all attempts to define word meanings by other means than corpus doomed to fail? Can translational equivalence be defined independent of the daily practise of translated texts? However interesting these questions are, they do not concern the set-up of a multilingual lexical database as such. Therefore, the discussion of these issues will be postponed until section 3.3.

There is a more directly relevant aspect of corpus based methods of MLLD's, which I will briefly discuss here. Multilingual computer corpus lexicography tries to give translations of words by means of corpus evidence of actual translations, found in parallel corpora. Parallel corpora consist of texts and their translation in another language, where the sentences (and possibly also smaller units) are explicitly aligned. The analysis of parallel corpora is extremely difficult, partly because translations often do not nicely line up with the original text because of structural differences between the languages, cultural dependency of the constructions used, or simply the artistic freedom of the translator. But I do not want to discuss the problems of parallel corpora in this thesis, but look beyond these 'practical' problems to the more fundamental issues.

What parallel corpora render is (ideally) pairs of words or phrases, of which the second has been (commonly) used as a translation for the first. Since these are relations between two languages, corpus methods fundamentally render pairwise linkings between languages. But they have a different way of reducing the work load of linking large numbers of languages: by performing the process of linking itself automatically. So in the light of multilingual computer corpus lexicography, do we still even need

multilingual lexical databases?

The reason that we do is that corpus methods, like the computer use approaches, have a different purpose and therefore different requirements from the kind of multilingual lexical database aimed for in the thesis. This difference becomes clear if we consider lexical gaps in computer corpus methods. As we have previously seen, in case of lexical gaps, one of the languages usually has a semantically more restricted term. Therefore, translations in case of a lexical gap are always incomplete translations. The projects discussed in this chapter so far try to model this explicitly by having special hyponym-translation relations. Corpus methods have no such possibility: something is either used as a translation or not; if this translation is incomplete, this does no longer show in the translation itself.

Take a concrete example: the English word 'finger' will normally be translated into Spanish with the word 'dedo'. So corpus evidence will show that a finger is a *dedo*, without any indication that it is in fact specifically a *dedo del mano*. So one could say that the MCCL approach only gets it right in the easy cases, where there is no lexical gap. But there is something to be said in favour of the corpus approach (Janssen, 2001): the difference between 'dedo' and 'dedo del mano' as a translation of 'finger' is the difference between what Zgusta (1971 [320]) calls the *translational equivalent* and the *explanatory equivalent*. Where no perfect translation of a word exists, there are two alternative ways of translating it: on the one hand a word (or a number of words) can be given that, although not a perfect translation, comes as close as possible and can be directly applied in the translation of a piece of text. This is the translational equivalent.

On the other hand, it is also possible to give an explanatory equivalent: an as good as possible explanation of the meaning of the word. Such an explanation will, however, in general not be directly applicable as a translation in a text. You could say that whereas the explanatory equivalent is more appropriate for a comprehension-oriented dictionary (meant for native speakers of the target language), the production-oriented dictionary user is better served by a translational equivalent. This important distinction by Zgusta directly applies to the translation generated by SIM*u*LLDA: even though 'dedo del mano' is probably the best explanation of the meaning of the word 'finger', it is not the expression that should be used when translating English into Spanish.

This thesis will opt for the comprehension-oriented method for two reasons: firstly, it gives a more precise model of the word meanings, since it gives more information on the more detailed meaning differences between languages. Secondly, because of the lack of differentiating information, it will never be possible to generate an explanatory equivalent from a translational one, whereas the reverse is less hopeless. The question how this could be done will be treated later on in this thesis. It is because of this emphasis on explanatory equivalents that corpus based methods do not serve

as a good basis for the SIMuLLDA system which will be explained in this thesis.

1.3 Conclusion to Chapter 1

In this chapter I have argued that a multilingual lexical database should meet a number of minimal requirements. I have done this by discussing difficulties that some of the most dominant and current multilingual lexicographic projects have and do not overcome. To list these requirements once again:

1. Languages will be connected at the level of meanings, and not at the level of words. Words can display ambiguities, that are arbitrary and by no means identical across languages. Hence, words of different languages are never said to be translationally identical, At best they share a meaning.

2. Since the number of language pairs in a database with an increasing number of languages grows exponentially, the meanings of the different languages should not be linked up pairwise, but are to be connected via an intermediate set of meanings. Since using one of the languages as such an intermediate structure would make the system dependent on the particularities of that central language, this intermediate structure should be a language-independent, interlingual set of meanings, to evade a number of undesirable effects.

3. Since not all interlingual meanings are lexicalised in every language, there will be lexical gaps. As we do not want these lexical gaps to result in the intranslatability of the related words, a mechanism should be present to produce translations in such cases. It would be unproductive and undesirable to resolve this problem by forcing every meaning in the interlingua to be expressed in every language, in order to avoid the existence of lexical gaps.

4. In order to meet the previous requirement, there has to be an ordering structure on the meanings of the interlingua. The differentiae specificae should be adopted into this ordering so as to be able to create proper explanatory equivalents.

The rest of this thesis will be devoted to describing a system that might fulfil these requirements. This system, which will be called SIMuLLDA, is not an existing system, but it is a theoretical framework. The purpose of this thesis is not only to describe the SIMuLLDA system, but also to test whether it does indeed meet the requirements set out above. SIMuLLDA is meant to be a lexicographer's tool, so one of the questions that should

be answered is whether system could indeed help in the actual creation of dictionaries, or only be a restricting burden.

The description of the system will involve the following steps: in chapter 2, the heart of the system will be introduced – a mathematical framework called Formal Concept Analysis, which provides the actual structure on the interlingua. Both the basic layout of FCA, and the way it is applied to lexicographic data in the SIM*u*LLDA set-up will be discussed. The basic layout in chapter 2 already provides the means to fulfil most of the requirements set out above, and especially the way lexical gaps are filled within the system will be discussed at length.

But for a more complete picture of the SIM*u*LLDA system, its basic elements and the claims behind it should be discussed in more detail. This will be done in chapter 3. The SIM*u*LLDA set-up has 4 basic types of elements: words, languages, interlingual meanings, and definitional attributes. These different elements and their precise nature will be discussed in turn. The reason behind this thorough discussion of the basic elements is the assumption that in order to avoid the many pitfalls associated with words and their meanings, one should take great care to provide all structure at precisely the correct level.

After the discussion of the basic layout, the system will be tested against some actual dictionary data. This will be done in chapter 4. In that chapter, I will show that the structure of the SIM*u*LLDA set-up is indeed rich enough to deal with an entire field of words from a number of languages, despite the many problems one encounters. There are even some problems that do lead to undesirable results, and restrictions on the possible content of dictionaries. But I will also try to show that these problems are unavoidable for any formal system dealing with dictionary data.

The system described in this thesis will not deal with all types of data that are present in dictionaries. The main discussion is restricted to the semantic characterisation of entity nouns. Also the empirical test in chapter 4 will be restricted to this type of dictionary data. Chapter 5 will discuss some of the dictionary data that are left out because of this restriction. Also, some aspects of the possible implementation of the system will be discussed: a mechanism for restricting the output of the system, and part of the actual programming. But first the next chapter will deal with the discussion of the FCA framework.

Chapter 2

FCA and SIM*u*LLDA

In the previous chapter, we have argued that a multilingual lexical database should ideally have a structured set of language-independent meanings, operating as its interlingua. The way the interlingual meanings are structured is the main issue of the system proposed in this thesis, which is called SIM*u*LLDA. In the SIM*u*LLDA system, this structure is provided by a simple and elegant, but also very rigid and powerful logical formalism called Formal Concept Analysis (FCA).

This chapter will start with an outline of the basic formalism of FCA. This basic formalism as such has nothing to do with dictionaries. However, the system can be applied to the content of dictionaries, as originally suggested by Uta Priß (1996). The SIM*u*LLDA system that will be outlined in this thesis uses a slightly different way of applying FCA to dictionaries, as will be outlined in section 2.3. After that, the system will be looked upon from a logical point of view (section 2.4), and partly implemented (section 2.5). But first, I will outline the basic principles of Formal Concept Analysis.

2.1 Formal Concept Analysis

Formal Concept Analysis (henceforth FCA) was developed by Ganter and Wille in Darmstadt (Ganter & Wille, 1996). It is an attempt to give a formal definition of the notion of a 'concept', within the boundaries of a model-theoretic framework.

Usually, research on concepts starts with an intuitive notion of existing, everyday concepts, and then tries to find characterisations of the objects belonging to that concept, for instance in terms of necessary and sufficient conditions (as will be discussed in section 3.3.2). FCA takes a different stance, and tries to give a formal notion of the nature of concepts, independent of any particular concepts.

FCA is a logical framework, which can be explained entirely in terms of

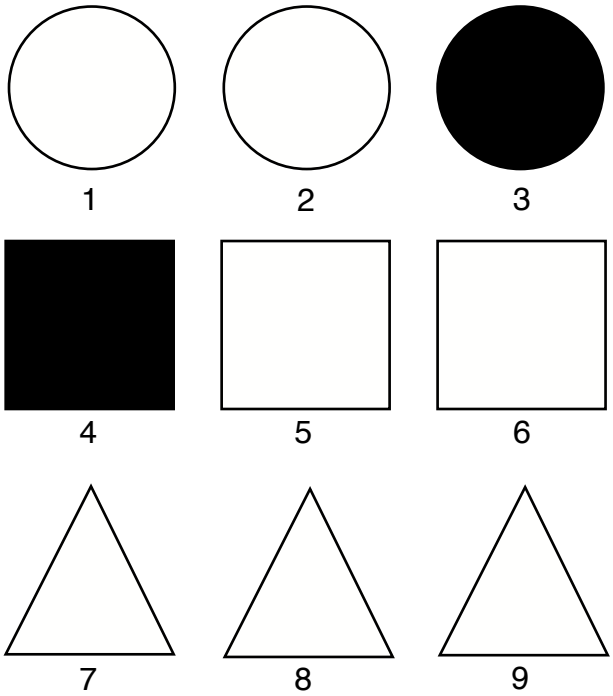


Figure 2.1: Very Simple Toy Model

abstract formulas, without any reference to the intuitions behind it. However, there is a very clear and simple intuition behind it, and with this intuition, FCA is not only a very elegant, but also a very natural way of defining concepts in a world model. Therefore, I will here try to explain the system in a very hands-on manner, using a very simple toy model.

Imagine that we have a 'world' consisting of only 9 objects. These objects themselves are extremely simple in terms of their properties: they only have a colour and a basic form. The objects of this toy model are represented in figure 2.1, where the objects are numbered 1-9.

Despite its simplicity and the relative randomness of the model, the objects in this toy model can be divided into naturally occurring classes. For instance, the objects 5 and 6 naturally belong together, and they do so for a very simple reason: they are (all) the objects that are both grey and square at the same time; they constitute the set of *grey squares*. It is very natural to view this grey-squaredness as a concept, and it is this notion of a concept that FCA tries to formalise.

If this is our notion of a concept, then concepts appear in every model, even in a simple and arbitrary model such as the one in figure 2.1. Concepts are related to sets of objects (2 and 5) and attributes (grey and square), but they are more than just arbitrary sets: 2 and 3 do not naturally form a

concept together, for although they are both round, they do not form the set of rounds, since 1 is not included. Also, concepts exist independent of our conceptualisation, and are independent of the kinds of objects and attributes involved.

So we can abstract away from the actual objects in the toy model, and represent the things as well as their attributes as abstract entities. If the objects are represented with the number $\{1 \ldots 9\}$ and the attributes (or features) that these objects have (round, square, triangular, white, grey, and black) with the two letter names in {ro, sq, tr, wh, gr, bl}, we can represent the toy model in figure 2.1 by table 2.1.

	ro	sq	tr	wh	gr	bl
1	×				×	
2	×			×		
3	×					×
4		×				×
5		×			×	
6		×			×	
7			×	×		
8			×	×		
9			×	×		

Table 2.1: Tabular Representation of the Toy Model

Within this abstract representation, we can say that the objects of the set {5,6} naturally belong together, because they share the attributes in the set {sq, gr}. This can be stated more precisely by saying that the pair $\langle\{5,6\},\{\text{sq, gr}\}\rangle$ constitutes what we call a *formal concept* (for which 'grey squares' is a useful, though arbitrary name), because all the objects in the set on the left of the pair have all the attributes in the right-hand set and conversely, all the attributes in the right set are shared by the objects in left set. The idea behind FCA is that concepts are precisely all the pairs of objects and attributes that have such a mutual dependency.

	sq	gr
5	×	×
6	×	×

Table 2.2: Sub-table for a Formal Concept

Ganter & Wille (1996) capture this idea in a formal logical framework, thus precisely defining the idea stated above. Formally, a context (the world) is defined as a set of objects G (Gegenstände), a set of attributes M

(Merkmalen), and a relation I between objects and attributes ($I \subseteq G \times M$), where $(g, m) \in I$ should be read as 'object g has attribute m'. We also define two functions: $^{\uparrow}$ and $^{\downarrow}$, where $^{\uparrow}$ is a function that takes objects and yields all the attributes that are shared by these objects, whereas $^{\downarrow}$ conversely takes attributes and yields the objects that share them. We define the set of formal concepts $\mathfrak{B}$ over a context (G, M, I) in the following way:

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

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

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

There is also a more visual way to see when a set of objects and a set of attributes (in a table) together form a formal concept: if we draw a sub-table for the objects and attributes, all squares have to be filled, as in table 2.2. And it should be impossible to find a bigger sub-table that is still completely filled[1]. So formal concepts are natural constellations in cross-tables. Thus we can conclude that every cross-table 'contains' formal concepts, even without a model like the one in figure 2.1 behind it.

Since $^{\uparrow}$ and $^{\downarrow}$ apply to a different domain, they are not easily confused. Therefore, given their similar role, it is often more convenient to use the same symbol for both of them. Following this standard convention in FCA, we will usually write A' for $A^{\uparrow}$ and B' for $B^{\downarrow}$. Notice that thus A'' is a function from a set of objects to a (possibly different) set of objects, namely those objects that have all the attributes that are shared by the objects in A.

Thus, a formal concept is a pair of a set of objects that have common attributes (the extent), and the defining set of attributes that they have in common (the intent)[2]. For convenience, for a concept $\alpha = \langle A, B\rangle$, we define two functions $\mathit{ext}(\alpha) = A$ and $\mathit{int}(\alpha) = B$. The concept is *realized* by the objects in its extent, and it is *defined* by the attributes in its intent.

Given these definition, we can determine the set $\mathfrak{B}(G, M, I)$ of formal concepts over our toy model. There are 13 concepts in total, and they are listed in table 2.3.

There are three things that should be noticed about this set of concepts. The first is the presence of the first concept 'Objects'. All the objects in figure 2.1 are objects, so the extent of this concept is $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. If we ask ourselves which attributes all these objects share, the answer is that they share *no attributes at all*. So the intent of this most general concept contains no attributes at all, otherwise put it is the *empty set* $\emptyset$. That this

[1] We will prove in section 2.4.6 that this informal characterisation is identical to the formal one in (2.3).

[2] These are technical notions: simply the left and right hand side of concepts. They remind of, but are not strictly related to the more philosophical notions of intension and extension.

Objects ($\top$)	$\langle\{1,2,3,4,5,6,7,8,9\},\emptyset\rangle$
Circles	$\langle\{1,2,3\},\{\text{ro}\}\rangle$
Squares	$\langle\{4,5,6\},\{\text{sq}\}\rangle$
Grey Object	$\langle\{1,5,6\},\{\text{gr}\}\rangle$
White Objects	$\langle\{2,7,8,9\},\{\text{wh}\}\rangle$
Black Objects	$\langle\{3,4\},\{\text{bl}\}\rangle$
Grey Circles	$\langle\{1\},\{\text{gr,ro}\}\rangle$
White Circles	$\langle\{2\},\{\text{wh,ro}\}\rangle$
Black Circles	$\langle\{3\},\{\text{bl,ro}\}\rangle$
Grey Squares	$\langle\{5,6\},\{\text{gr,sq}\}\rangle$
Black Squares	$\langle\{4\},\{\text{bl,sq}\}\rangle$
White Triangles	$\langle\{7,8,9\},\{\text{wh,tr}\}\rangle$
$\perp$	$\langle\emptyset,\{\text{ro,sq,tr,gr,wh,bl}\}\rangle$

Table 2.3: Formal Concepts of the Toy Model

most general concept, usually called *top* or $\top$, is a formal concept according to the definition above is easy to see: if there are no attributes, then these non-existing attributes are trivially shared by all the objects, and there is no object that doesn't share one of those non-existing attributes.

The second thing to notice is the last concept in the list, the least general concept, called *bottom* or $\perp$. It is the inverse of the top: it is the concept defined by all the attributes at the same time. But since there are no objects that are (for instance) both black and white at the same time, the extent of this concept will be empty. For the record: there *can* be objects in the extent of $\perp$, if there are objects in the context that have all the available attributes, and there also *can* be attributes in the intent of top, for instance when we add 'flat' as an attribute to all the objects in the model.

The last thing to notice is the absence of the concept 'Triangle'. The reason why there is no such concept is that all the triangles in our toy model happen to be also white. That means that what one would expect to be the concept 'Triangle' (that is $\langle\{7,8,9\},\{\text{tr}\}\rangle$) is not a formal concept, since the right-hand side does not contain all the attributes shared by the left hand side; the projection $\{7,8,9\}'$ is $\{\text{tr, wh}\}$. So within the framework of FCA, we have only the more specific concept 'White Triangle'. Why this turns out to be a desirable property, we will see later on.

2.1.1 Partial Ordering

Not only is there a natural set of concepts for an arbitrary model (context) such as the one in figure 2.1; these concepts also have a natural order amongst themselves. Reconsider the concept of 'Grey Squares'. The objects

within the extent of this concept are much like the objects in the extent of the concept 'black squares' $\langle\{4\}, \{\text{sq, bl}\}\rangle$ in that they are all squares. This can also be stated in the following way: there is a more general concept 'Squares' $\langle\{4, 5, 6\}, \{\text{sq}\}\rangle$, which contains the concepts 'Grey Squares' and 'Black Squares' as specific instances.

We know that 'grey squares' is what you might call a *subconcept* of 'squares' for two reasons:

1. all the 'grey squares' are also 'squares'

2. all the attributes that define the 'squares' also define the 'grey squares'

Given the definition of formal concepts in FCA, these two statements are actually different ways of saying exactly the same.

To formalise the notion of subconcept: a concept α is a subconcept of a concept β, if and only if α has all the defining attributes β has and possibly some more. This is the same as saying that a concept α is a subconcept of a concept β, if and only if all the objects in the extent of α are also in the extent of β:

(2.4) $\alpha \leq \beta \Leftrightarrow \mathit{ext}(\alpha) \subseteq \mathit{ext}(\beta) \Leftrightarrow \mathit{int}(\beta) \subseteq \mathit{int}(\alpha)$

This subconcept-relations puts order on the concepts. It is not a strict ordering, but a *partial ordering*. If you look at things that are strictly ordered, like the natural numbers, it holds that if you take two arbitrary objects (numbers), the first is always either smaller or bigger than the other (or identical if you allow to take the same object twice):

(2.5) $\forall x.\forall y.x \neq y \Rightarrow x < y \vee x > y$

For concepts this clearly doesn't hold: 'grey squares' is neither a sub- nor a superconcept of 'black squares' though both are subconcepts of 'squares'. Therefore it is merely a partial ordering. Also, concepts do not form a hierarchical tree: 'grey squares' is a subconcept of 'squares'. But in the same right, 'grey squares' is also a subconcept of 'grey objects'. And in a normal hierarchy, concepts would have only one superconcept. So it is a kind of hierarchy, but one that is usually referred to as a *multiple inheritance network*.

Although not every set of objects constitutes the extent of a concept, every set of objects A can be projected onto what you might call its *smallest common concept*: the smallest concept to which all the objects in A belong. The way to get this smallest common concept is simple: take all the attributes they have in common (which is by definition the set A'), and then take all the objects that share these attributes, or $(A')'$; so take the pair

$\langle A'', A' \rangle$. In order to show that $\langle A'', A' \rangle$ is indeed the smallest common concept we need to show that this pair is indeed a concept (2.6), and that it is also the smallest concept (2.7):

(2.6) $\forall A \subseteq G. \langle A'', A' \rangle \in \mathfrak{B}(G, M, I)$

(2.7) $\forall A \subseteq G. \forall \alpha \in \mathfrak{B}(G, M, I). A \subseteq \text{ext}(\alpha) \Rightarrow \alpha \geq \langle A'', A' \rangle$

For (2.6) we need to show that $A = B'$ and $A' = B$. The first is immediately given: $(A')' = A''$, and the second is a direct result of a standard rule in FCA: $A' = A'''$ (see section 2.4.1). Equation (2.7) is slightly more difficult to prove and will be proven in section 2.4.2.

To give an example of a smallest common concept: for the set {2,3}, the only attribute they have in common (being $\{2, 3\}'$) is {ro}, and the set of all round objects (or {ro}$'$) is {1,2,3}. So the smallest common concept $\langle \{2, 3\}'', \{2, 3\}' \rangle$ is the concept 'Circles': $\langle \{1, 2, 3\}, \{\text{ro}\} \rangle$. Of course we can not only take the smallest common concept for sets containing many objects, but also for sets containing just one object. So every object has a 'smallest common concept' to which it has a special relation; for instance, for object 1, this is $\langle \{1\}'', \{1\}' \rangle = \langle \{1\}, \{\text{gr, ro}\} \rangle$, or the concept of 'Grey Circles'. Since the formal concepts themselves have no names, it is sometimes useful to use the formal object as a way of indicating its smallest common concept.

Given the special concepts $\perp$ and $\top$, the order on concepts has the following property: for every arbitrary set of concepts, there will always be a concept that is the smallest superconcept of all of them ($\top$ is a superconcept of everything), and there is also a concept that is the largest subconcept of them all ($\perp$ is a subconcept of everything). A partial ordering with this additional property is called a *(complete) lattice*[3].

We can now deduce the ordered set of formal concepts for the toy model in figure 2.1. For the 13 concepts in table 2.3, the following intuitive nomenclature will be used: W for the white objects, S for the squares, GC for the grey circles, etc. Using this notation, the 13 concepts are:

(2.8) $\mathfrak{B} = \{\top, W, G, B, C, S, WT, WC, GC, BC, GS, BS, \perp\}$

The ordering can be easily read from this notation: the concepts with two letters are subconcepts of the one-letter concepts with the two constituting letters. So $WC \leq W$ and $WC \leq C$, etc.

Note that the number of formal concepts is not a straightforward result of the fact that there are 9 formal objects and 6 formal attributes. Given the number of objects and attributes, there will be a maximum number of formal concepts ($2^6 = 64$, which can be reduced to 17, as will be explained in

[3]When only pairs of concepts are considered, it is just a lattice; for finite cases, all lattices are complete.

section 2.4.6) but the specific number of formal concepts for a given context depends upon the actual model itself; for instance, the fact that there is no non-white triangle, makes that there is no specific concept *triangle*, which could not have been predicted by only looking at the objects and the attributes[4]. So the fact that this model has 13 formal concepts is more or less accidental.

2.1.2 Hasse Diagrams

The ordered set of 13 formal concepts resulting from the toy model in figure 2.1 can be graphically represented using the following convention: if a concept α is a subconcept of a concept β (*smaller* regarding to the partial ordering relation $\leq$), we put α *below* β, and connect them with a line. A graphical representation of a partial ordering using this convention is called a *Hasse-diagram*. Note that the possibility to draw a Hasse-diagram depends on the fact that if a concept α is 'smaller' than a concept β, it can never at the same time be 'bigger' than β. This property is called *anti-symmetry*: $\alpha \leq \beta \ \& \ \alpha \geq \beta \Rightarrow \alpha = \beta$.

The Hasse-diagram for the simple toy model in figure 2.1 is given in figure 2.2. With the 13 different concepts over the 9 objects in the domain, it has a more informative structure than the model itself. The horizontal positioning in Hasse diagrams is arbitrary; more on this will be said in section 2.5.

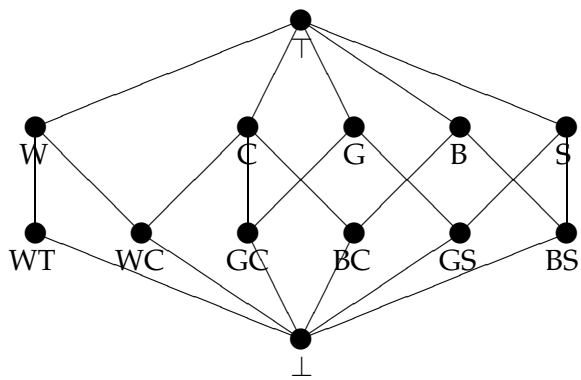


Figure 2.2: Concept Lattice of the Toy Model

Notice that the relation between the flat information in table 2.1 and the Hasse-diagram in figure 2.2 is completely independent of the interpretation of the rows and columns of the table; the only thing that is relevant for the

[4]Of course it could have been predicted if you also looked at the relation I, but then you would have taken the entire model into consideration.

Hasse-diagram is the distribution of the crosses in the table. Therefore, any table can be graphically represented using FCA.

Within the tradition of FCA, there is an additional convention for drawing Hasse-diagrams: for formal concepts, it is sometimes more instructive to represent not so much the arbitrary names given to the concepts, but their intent and their extent in the Hasse-diagram (although sometimes their names are more clear). As a convention, the objects of the extent are written *below* the node representing the concept, whereas the attributes of the intent of the concept are written *above* the node.

If there are many objects and attributes in the context, this will result in a rather illegible diagram, containing much superfluous information. There is superfluous information because if there is an object in the extent of a concept, it will by definition also be an element of the extent of all its superconcepts; and since all superconcepts are represented directly above their subconcept, connected by a line, and hence easily traceable, the items in the extent need only be represented on the most specific (furthest down in the lattice) concept they appear in. So formal objects are represented exactly once in the Hasse-diagram, below the node of the most specific concept for which they are in the intent. This most general node is of course the smallest common concept of the object, which by equation (2.6) is the concept $\langle g'', g' \rangle$[5] for any $g \in G$.

Because the system of FCA is completely symmetrical with respect to objects and attributes, we can do the same for attributes: attributes need only be represented above the node of the most general concept that they are part of the intent of, being the concept $\langle m', m'' \rangle$ for any $m \in M$.

If we use this convention on the lattice for our toy model, we get the diagram represented in figure 2.3. Often you will see mixed versions of these two systems, where both concept labels *and* intent and extent are represented. This of course will be redundant information, but helps to make the lattice intelligible.

2.2 Connotative Context

FCA is a formal framework, which means that although the two components of formal contexts are called 'objects' and 'attributes' respectively, they are both no more than abstract entities. Also the notion of a 'concept' that is defined by FCA has as such nothing to do with concepts. Formal concepts are no more than what they are defined to be: pairs of formal objects and formal attributes, with a relation between them defined by the functions $^{\uparrow}$ and $^{\downarrow}$.

The system acquires a meaning only by giving an interpretation to these entities; making these entities stand for something. The most natural inter-

[5] Actually $\langle \{g\}', \{g\}'' \rangle$, but we will write g' as well as $\{g\}'$ for singletons.

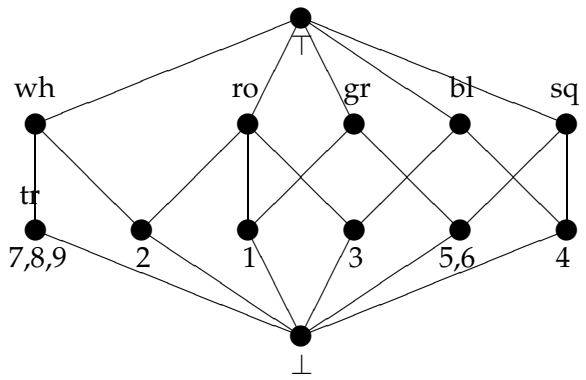


Figure 2.3: Concept Lattice with Intent and Extent

pretation of Formal Concept Analysis is to interpret the formal objects as representing real objects in the external world, and the formal attributes as representing the properties that these objects have. But that is by no means the only possible interpretation. Another possible interpretation (however pointless) is the following: let the formal attributes stand for objects in the world, and the formal objects for the properties that they have. In other words, interchange the interpretation of objects and attributes. That would lead to formal concepts where the extent of the formal concept is the intention of the real world concept; in which the sub-concept relation indicates a real-world super-concept, etc. The resulting system would be almost identical to the more natural one (isomorphic when the order is reversed), since in FCA objects and attributes are defined completely symmetrical. The only difference would be that all the names of the formal properties would be confusing at least.

A more interesting alternative interpretation of FCA, and the one that this thesis will focus on, is one relating to words and word meanings. This interpretation of FCA was originally introduced by Priß (1996) in her thesis, as part of a larger system. Priß calls the contexts that have their natural interpretation (where the formal objects are interpreted as real objects, and the formal attributes as real attributes) *denotative contexts*. She also describes a different kind of context, which she calls *connotative* contexts. In a connotative context, the formal objects are *word meanings*, and the formal attributes are the attributes related to these word meanings:

> A *connotative context* $\mathcal{K}_K := (M(W), A_K, I_K)$ is defined as a formal context whose formal objects are particular meanings and whose formal attributes are features of the particular meanings. The set of particular meanings is denoted by $M(W)$, the set of features of the particular meanings by A_K, and a relation that assigns features to particular meanings by I_K. The concept lattic $(\mathcal{K}_K)$ of a connotative context $\mathcal{K}_K$

is called *connotative lattice*. (Priß, 1996 [23])

Connotative and denotative contexts are linked in the system proposed by Priß, as illustrated in figure 2.4. On top of the denotative structure and the connotative structure (both of which are defined in terms of FCA), there is also a lexical structure, in which the different aspects of the word itself are modelled. This lexical structure is not defined in term of FCA.

The structure in figure 2.4 explicitly models the idea by Gottlob Frege (1892) that there are two sides to the meaning of a word: the *Sinn* (commonly translated as sense or connotation) and the *Bedeutung* (commonly translated as meaning or denotation). The denotation (of a noun) is the set of objects in the world denoted by the word, whereas the connotation is the way in which this object is denoted. His standard example is that although the words *Hesperus* and *Phosphorus* denote the same object (the planet Venus), they are still different words, in/with a different sense.

The lexical semantic theory that Priß proposes is a very rich system, modelling many aspects of the meaning and structure of words. It describes these aspects in a very strict and precise way, which has both an advantage and a disadvantage: on the positive side, it very explicitly describes all the relations that are assumed to exist between the various components of the theory, hence making many implicit aspects of many lexical theories explicit.

But on the negative side, this also makes it very clear where the theory assumes too much structure on the world and on words. By the very fact that there is a denotative context, the theory assumes the world to be nicely ordered in objects and attributes. Also, by the fact that the denotative word concepts form a subset of the denotative concepts, it assumes our lexicon to correctly follow this inherent structure of the world. Since there is a mapping (dnt) from the connotative word concepts to denotative word concepts, the intentional meanings are assumed to uniquely determine their denotation. And all the other mappings create similar predictions.

If there is one thing that lexical semantic theories have shown, it is that (almost) no relation between words and anything else is completely tenable. Some of these issues will be discussed in the next chapter; for instance, the shortcomings of denotational semantics will be discussed in section 3.3.2. Given these limitations of lexical semantics, the proposal in this thesis will take an almost opposite position from the one presented by Uta Priß: where figure 2.4 takes a very rich theory, with much assumptions on the structure of the world, SIMuLLDA will take a very shallow theory of word meaning, where as little structure on words is presupposed as possible.

So though the system proposed in this system is based upon upon the same idea as the proposal by Priß (apply FCA to word-meanings and their

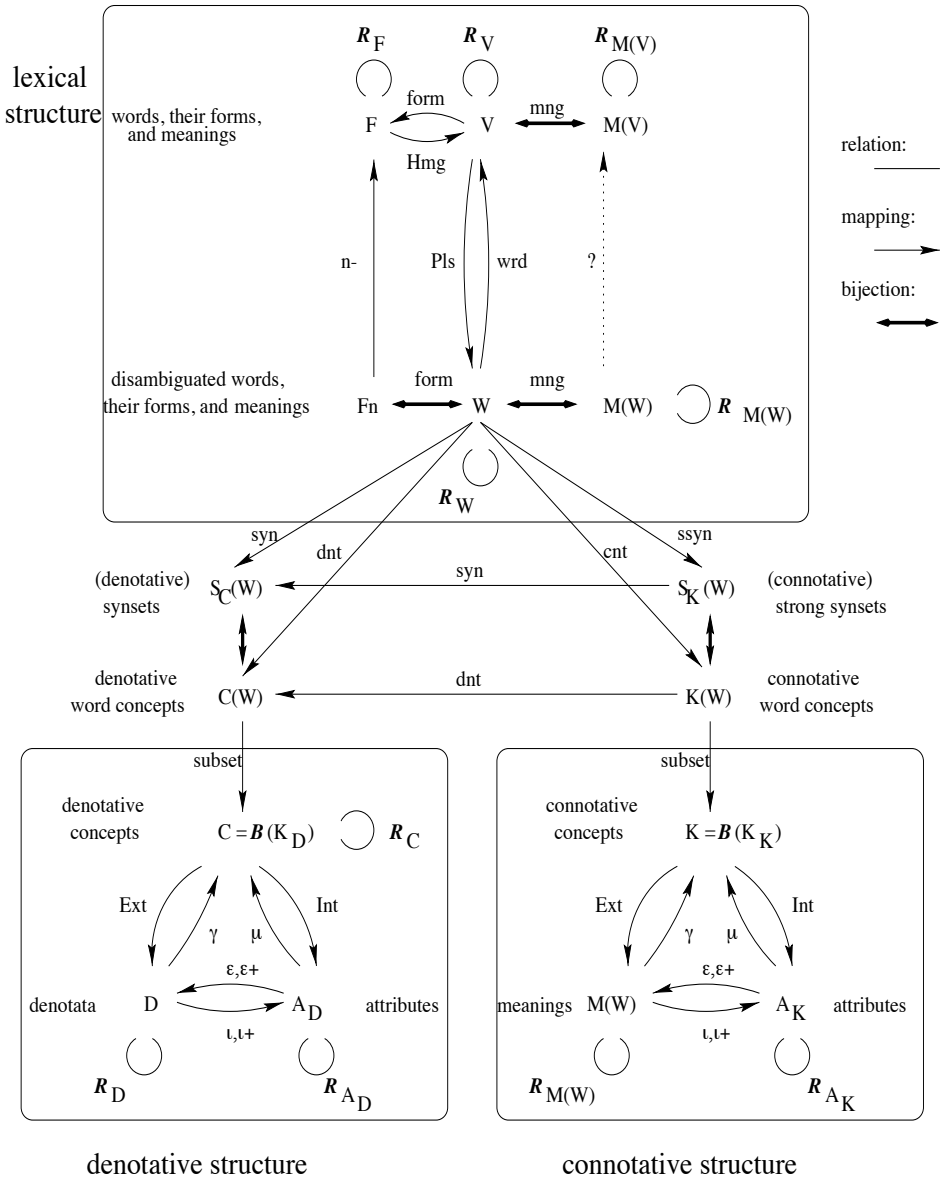


Figure 2.4: Connotative Structures in Context (Priß, 1996 [21])

attributes), the way this idea is worked out is largely different, especially in its details. As a consequence, no detailed analysis of the proposal by Priß will be given; but let me give an idea of what the basic elements of connotative contexts are. According to Priß, the connotative context is closely related to the linguistic notion of word meaning: *"Connotative concepts often represent what linguists (for example Saussure (1972)) mean when they say the meaning of a word is a concept."* (Priß, 1996 [22]). The definitional attributes in connotative context are all those aspects of word senses that are not linked to the denotation. This involves on the one hand those things that would belong to the Sinn of the word: *seen in the morning* for *Hesperus* and *seen in the evening* for *Phosphorus* (Priß, 1996 [32]). But on the other hand it also involves the more pragmatic aspects of words, such as *common language* for *common dog*, vs. *biological term* for *Canis familiaris* (Priß, 1996 [24]).

Connotative contexts are explicitly designed to be able to deal with lexicographic data. In her thesis, Uta Priß discusses how three types of lexicographic data could be interpreted in terms of FCA with connotative contexts: the Webster Third, the Roget's International Thesaurus, and WordNet. Also, Uta Priß has used an extension of the theory (called Relational Concept Analysis) as a tool to graphically display the content of WordNet, so that inconsistencies in the relations became more visible.

Despite this applicability to dictionaries, connotative contexts do not directly lead to a multilingual lexical database as described in the previous chapter. The main reason for this is that the definitional attributes of connotative contexts are not worked out in sufficient detail for an analysis of the actual definitions in dictionaries to lead to the kind of structure that was argued for in the previous chapter.

This thesis will present a different approach to FCA in combination with words and lexical definitions, that hopefully does provide such a structure. This alternative approach will be the heart of the SIM*u*LLDA system. We will not start from the elaborate structure in figure 2.4, but build the system up from scratch, and try to come to as simple and elegant as possible a system where FCA is applied to dictionary definitions.

2.3 The SIMuLLDA System

The idea behind SIM*u*LLDA is in a way similar to that behind connotative context: apply Formal Concept Analysis to word-meanings and their attributes. However, the kind of attributes that will be linked to word-meanings are different in SIM*u*LLDA. The idea is that they are closer to the semantic part of dictionary definitions. To distinguish them from connotative context, the contexts in SIM*u*LLDA will be called *lexicographic contexts*. Their nature is best shown using some actual dictionary definitions,

such as the definitions of English words for kinds of horses as found in the Longman Dictionary of Contemporary English (LDOCE). The definitions are displayed in table 2.4.

colt	/kəʊlt/ *n* a young male horse – compare FILLY
fil·ly	/ˈfɪli/ *n* a young female horse – compare COLT
foal[1]	/fəʊl/ *n* a young horse
mare	/meəʳ/ *n* a female horse or DONKEY – compare STALLION
stal·lion	/ˈstæljən/ *n* a fully-grown male horse kept for breeding – compare MARE

Table 2.4: Definitions in LDOCE for Horses

These dictionary definitions can be straightforwardly interpreted as describing an FCA context, in the following way: the left-hand side of the definition already contains words and hence can be simply taken as formal objects for lexicographic contexts. They are not words but rather disambiguated words (word-forms; their exact status will be discussed in the next chapter), since only one of their meanings is given: *colt* as referring to a kind of gun will be a different formal object in the lexicographic context.

The right-hand side of the definition can be easily interpreted as describing features of these word meanings. Take for instance the definition of *filly*: it says that the meaning of *filly* is the same as the meaning of *horse*, except that it is more restrictive in two respects: fillies are young, and fillies are female. So the differentiae specificae 'young' and 'female' are features, distinguishing the meaning of *filly* from the meaning of other hyponyms of *horse*. We will call these features *definitional attributes*, and those definitional attributes will be the formal attributes of lexicographic contexts.

But the differentiae do not form the entire definition of *filly*: there also is the genus proximum *horse*. This genus is again a disambiguated word; it relates to a specific meaning of the word *horse*, since it is not supposed to refer to 'a kind of gymnastic equipment'. This word meaning again has a definition in the dictionary, consisting of genus proximum et differentiae specificae. So we can further 'unfold' the definition of *colt* into more definitional attributes with a new genus term. The idea behind SIMuLLDA is, that if you unfold the definitions in this way, you completely reduce the dictionary definition of a word to a set of definitional attributes[6]. There are, of course, many problems related to this process of unfolding: not all definitions have the form of genus et differentiae, some definitions are cyclic, especially with very general terms like 'thing' and 'object', and there has to be a termination point where the unfolding stops. These problems will be

[6]This is not a new idea; for instance, Vossen & Copestake also observe: *"[I]t is possible to trace the entry word/genus relation in dictionaries by recursively looking up the genus in the same source … thus for* moussaka *we not only can infer that it is 'Greek', 'made from meat and aubergines', and 'often has cheese on top', but also via* dish[1] 2 *that it is 'cooked', via* food *that it is 'eatable', and via* substance *that it is 'material'.* (Vossen & Copestake, 1993 [246])

discussed later on in this thesis. But in our current example it works nicely.

For the sake of the example, we will make a few simplifications to the definitions in table 2.4. Firstly, the genus term *horse* will not be treated as a genus term, but as most specific term: a word without a genus proximum, but with only a differentiam specificam. Such a 'top-word' is necessary, since otherwise the process of unfolding would never stop; in a complete system, this would be an empty word like *entity* or *thing*, but here we will take *horse* to only have a differentiam: 'horse'.

Also, the definitions of *mare* and *stallion* will be slightly modified for simplicity: firstly, the parts 'or DONKEY' and 'kept for breeding' are dropped. Secondly, the asymmetry that stallions have to be fully-grown, whereas mares do not is resolved, by opting that both should be adult horses. We will get back to these definitions without simplifications in chapter 4.

In this way, we can reduce the definitions in table 2.4 to simple sets of word meanings and definitional attributes. The result of this is given in table 2.5.

	horse	male	female	adult	young
horse	×				
stallion	×	×		×	
mare	×		×	×	
foal	×				×
filly	×		×		×
colt	×	×			×

Table 2.5: Analysis of Definitions for Horses

If we view this as a lexicographic context, we will get formal concepts related to these definitions, resulting in a lattice of the ordered set of lexicographic concepts. This lattice can be graphically displayed as in figure 2.5. Following convention, the labels under the nodes represent the formal objects (word meanings) at the lowest node where they appear in the extent of the related concept, while the labels above the nodes represent the formal attributes (definitional attributes) at the highest node where they appear in the intent.

In building a lattice out of the definitions in table 2.4, we have gained additional structure: we now have the explicit information that *mare* and *stallion* are related terms. There is even new information: whereas in the definitions it is only said that a *colt* is a young male horse, we now also have the information that it is a young foal. So at least in a monolingual setting, with simple examples, FCA yields a nice structure for dictionary definitions.

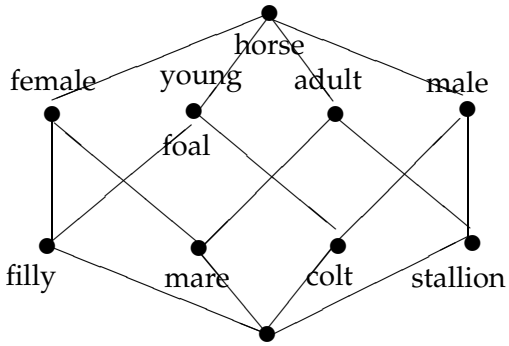


Figure 2.5: Concept Lattice for Horses

2.3.1 Multilinguality

Since this is a thesis on a multilingual lexical database, we want to use lexicographic contexts in a multilingual setting. In principle, this can be done in a very simple and straightforward way: in table 2.5, you find what the definition of *stallion* is: male adult horse. Now (virtually) every language has a word for a male adult horse, as well as for most of the other meanings in table 2.5. A small sample of the different words in some language is listed in table 2.6.

English	horse	stallion	mare	foal	filly	colt
Dutch	paard	hengst	merrie	veulen	merrieveulen	hengstveulen
German	Pferd	Hengst	Stute	Fohlen Füllen	Stutenfohlen Stutenfüllen	Hengstfohlen Hengstfüllen
French	cheval	étalon	jument	poulain	pouliche	*no spec. word*
Italian	cavallo	stallone	cavalla giumenta	puledro	puledra	*no spec. word*
Hungarian	ló	csõdör	kanca	csikó	fruska	*no spec. word*
Swedish	häst	hingst	sto	föl	ungsto	unghingst
Russian	ло́щадь	жеребе́ц	кобыла	жеребёнок	кобылица	*no spec. word*
Georgian	ცხენი		ჯიგი	ქვრივი	ფაცხენი	*no spec. word*
Malay	kuda	kuda jantan	kuda betina	anak kuda	anak kuda betina	anak kuda jantan

Table 2.6: Words for Horses in Different Languages

Since the words *hengst*, *Hengst*, *étalon*, *stallone*, etc. in the second column are all supposed to have the same meaning as *stallion*, and the meaning

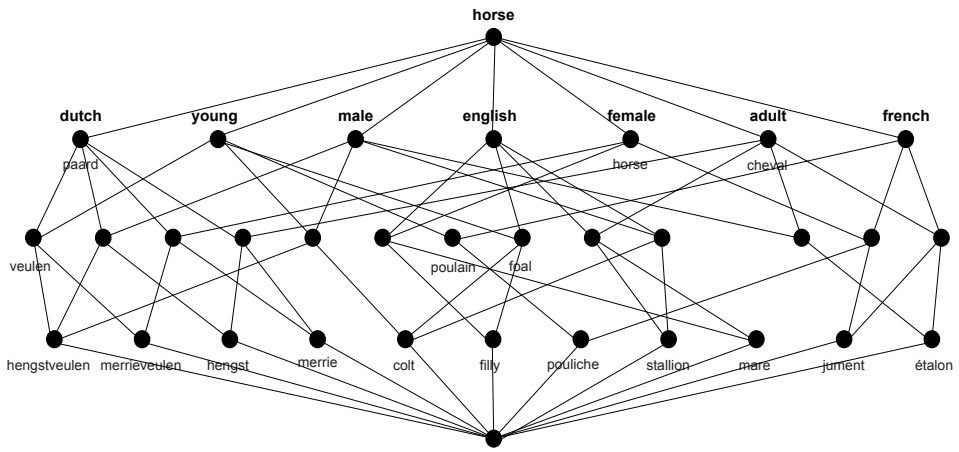


Figure 2.6: Multilingual Connotative Context

of *stallion* is defined in terms of the definitional attributes *horse*, *male*, and *adult*, these same definitional attributes should also define the meaning of *hengst*, *Hengst*, *étalon*, etc. And in the same way all the other words in table 2.6 should be reducible to definitional attributes. This will lead to a multilingual lexicographic context, in which all the words in table 2.6 are related to the five definitional attributes in table 2.5. And that in turn yields a multilingual concept lattice.

This way, it is even simple to predict what this interlingual concept lattice will look like: it will look exactly like the monolingual one in figure 2.5, except that the node that now just has the word *stallion* below it, will get *stallion* and all its translational synonyms below it. The reason is that formal objects that have exactly the same formal attributes will always appear together in every extent; compare the coinciding of the identical objects 5 and 6 (and 7, 8, and 9) in figure 2.3.

With this method, the words *stallion* and *hengst* would be indistinguishable; but there is an important difference between them: they belong to different languages. This information could easily be resolved by adding language as an additional definitional attribute. This would lead to the much more complex concept lattice in figure 2.6[7].

In principle, the network in figure 2.6 is a nice, interlingual concept lattice in which in principle all the words (meanings) of a language are nicely recognisable; the words of the French language are all the sub-nodes of the uppermost node with has **french** in its intent, i.e. $\langle \mathbf{french}', \mathbf{french}'' \rangle$.

[7]Notice that in this lattice, there are many nodes without a lexicalisation, such as the node directly above *colt* and *stallion*, representing the concept of 'English words for male horses'.

However, there is a problem with this representation: each word will appear in the lattice as many times as it has distinct meanings. And there is no direct access to the word itself, so the entire lattice has to be searched in order to get all the different meanings for a single word. Therefore, it would be very convenient to have a single entry for a word that directly gives access to all its various meanings[8].

The reason for that is rather simple: the lattice in fact does not contain words, but only word-meanings. So the reason why it is not a good multilingual lexical database is because the basic element of dictionaries, (the words) are absent from the system. Also, it is not an interlingual lattice, since the lattice explicitly contains elements that are language-dependent: the word-meanings of the words in the different languages. Therefore, this thesis will opt for a slightly different approach, and remove all language-dependent elements from the central system and into language-dependent structures, where they are linked to words. In that way, the lattice can truly operate as an interlingua, linking the various languages.

Language Boxes

In the SIM*u*LLDA set-up, the FCA lattice will be made language-independent by moving the word-forms themselves away from the lexicographic context into language boxes. Thus removing the word-forms, the formal objects in lexicographic contexts will no longer be language-dependent disambiguated words, but rather language-independent meanings, which are related to, but not identified with, the word-forms of the various languages. The word-forms themselves are situated outside of the lattice, and related to the meanings in the sense that every word-form expresses one or more meanings.

The word-forms are grouped into languages, where languages are taken as little more than lists of word-forms. This gives us a set-up as exemplified in figure 2.7. In this figure, the boxes in the figure represent the languages, containing word-forms. These word-forms refer to language-independent meanings in the interlingua, and this relation is indicated with a grey line. The language-independent meanings are the formal objects in the FCA context, and hence by convention represented below the lowest formal concept in the extent of which they appear. Above the nodes are the definitional attributes; the context producing the lattice is only implicitly present in this figure.

The definitional attributes in the figure are not (yet) linked to the different languages. Definitional attributes are, in a way, less obviously language-

[8]This is, of course, not a result of the multilinguality, but already present in the monolingual lexicographic contexts.

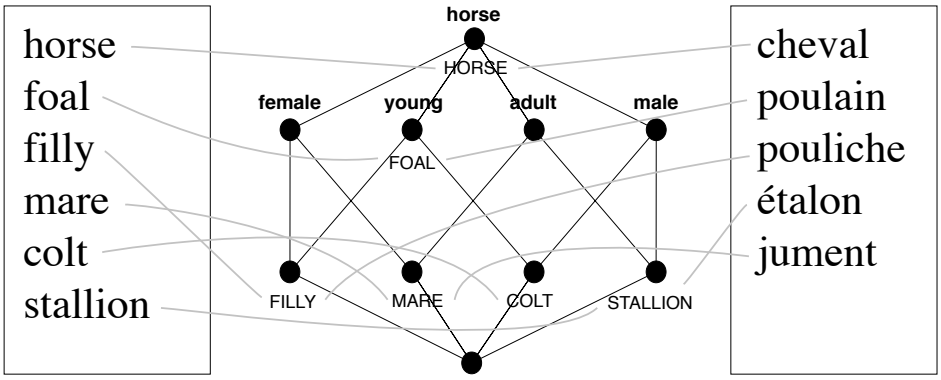


Figure 2.7: Partial Multilingual Set-up

dependent than words: the definitional attribute **young** expresses, in combination with the meaning FOAL, merely the fact that foals are young. That fact itself is not language related (what exactly definitional attributes are will be discussed in section 3.4). However, *young* itself is, of course, an English expression for this definitional attribute, which could equally well have been expressed by *jeune* in French, or *mlády* in Czech. So in the SIM*u*LLDA set-up, also the definitional attributes are related to expressions of the various languages, leaving the definitional attributes themselves completely language-independent.

In order to keep the different entities in SIM*u*LLDA apart, the following convention will be used: interlingual meanings will be indicated with SMALLCAPS, attributes will be indicated with **boldface**; the 'words' are not fully specified yet, and will for the time being be indicated with roman letters (all notational conventions used in this thesis can be found in appendix B.4). Of course, the language-independent items themselves (meanings and definitional attributes), do not have a written form. For clarity, they will be given the names of their English lexicalisations, indexed with a number where necessary. This naming is arbitrary, and has some exceptions: when various English lexicalisations exist (in case of synonymy), either of these is chosen, and when no English lexicalisation exists (in case of a lexical gap), the lexicalisation in an arbitrary other language will be chosen. We also want to be able to refer to the node (= formal concept) under which the word is represented, which is, as shown in the previous chapter, the smallest common concept. We will refer to this formal concept in bold small caps, so the node for the smallest common concept of the interlingual meaning COLT is $\langle \text{COLT}'', \text{COLT}' \rangle$, for which the shorthand **COLT** will be used. Logically, it holds that ext(**COLT**) $\supseteq$ COLT.

(2.9) $\textbf{COLT} := \langle \text{COLT}'', \text{COLT}' \rangle$

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

	English	French	Dutch
male	male	mâle	mannelijk
female	female	femelle	vrouwelijk
adult	adult	adulte	volwassen
young	young	jeune	jong
HORSE	horse	cheval	paard
STALL.	stallion	étalon	hengst
MARE	mare	jument	merrie
FOAL	foal	poulain	veulen
FILLY	filly	pouliche	merrieveulen
COLT	colt		hengstveulen

Table 2.7: Set-up for Horses in SIM*u*LLDA

Using these conventions, the multilingual set-up of the SIM*u*LLDA system (preliminary) is as represented in table 2.7: the language-independent meanings and definitional attributes constitute, together with the relation between them, the lexicographic context as depicted in the upper half of the table; the interlingual items can be expressed by word-forms in the different languages, as depicted in the lower half of the table. As discussed earlier, not all meanings have to be lexicalised in every language; there may be lexical gaps such as the lexical gap in French for *colt*. Definitional attributes, however, will be required to be lexicalised in every language. The reason for this will be explained in the next section, since it involves the process of lexical gap filling.

Since the relation between words and interlingual meanings is an important one, two functions will be introduced: *wfs* yields the set of word-forms related to a given interlingual meaning, and *mng* gives the interlingual meanings expressed by a 'word'. Where necessary, *wfs* will be indexed with the language in which the word-forms should be given. So in our example, $\text{mng}(\textsf{horse}) = \{\text{HORSE}\}$, and $\text{wfs}_{\text{English}}(\text{HORSE}) = \{\textsf{horse}\}$[9].

[9]The function wfs yield a set of synonymous word-form, resembling the WordNet synsets. So you could say that SIM*u*LLDA works with synsets like WordNet does, even

2.3.2 Lexical Gap Filling

One of the attractive features of the SIM*u*LLDA set-up is its capacity to construe translations for lexical gaps. How this works can be explained using the layout of SIM*u*LLDA in figure 2.7[10]. Normally, the translation of a word is rendered by SIM*u*LLDA in the following way: the word *horse* in the English language module expresses the interlingual meaning HORSE, as graphically indicated by the line connecting the two. This interlingual meaning can be lexicalised in French, since the meaning HORSE is related to the word *cheval* in the French language module. This yields a direct translational relation between the English word *horse* and the French word *cheval*. So *cheval* is a translational synonym of *horse* because $\text{wfs}_{\text{French}}(\text{mng}(\textsf{horse})) \supseteq \{\textsf{cheval}\}$[11].

For the English word *colt* however, this does not work: the meaning expressed by *colt* (COLT), has no lexicalisation in French. This implies that there is a lexical gap in French for the English word *colt*. So a lexical gap can be defined in the following way:

Lexical Gap There is a lexical gap in a languages Y for the word x in language X, if (one of) the interlingual meaning(s) expressed by x does not have a lexicalisation in language Y. In terms of the functions *wfs* and *mng*: if $\text{wfs}_Y(\text{mng}(x_X)) = \emptyset$.

However, because of the structure of the SIM*u*LLDA interlingua, even meanings that are not lexicalised in a particular language, can be translated into that language by the following mechanism: find a superconcept of the concept that is lexicalised in that language, and then look at the difference between the set of definitional attributes for this superconcept and the definitional attributes of the concept is to be translated. By definition, there will be a definitional surplus for the subconcept, which can be calculated in the following way:

Definitional Surplus For two concepts $\alpha = \langle A_1, B_1 \rangle$ and $\beta = \langle A_2, B_2 \rangle$, where $\alpha \leq \beta$, we define $\text{defsur}(\alpha, \beta) = B_1 \backslash B_2$

So the definitional surplus consists of the definitional attributes of the node itself, minus the definitional attributes of the superconcept. This definitional surplus, together with the lexicalisation of the superconcept, will give a complete translation of the meaning of the concept in the desired target language.

though these are only indirectly present, and represented by their interlingual meaning.

[10]The nature of words and languages still has to be established in the next chapter, which means that the description is incomplete at some points.

[11]The $\supseteq$ relation leaves open the possibility that there are other synonyms for *cheval* in French.

In the case of *colt* this works as follows: the English word *colt* expresses the meaning COLT, which is not lexicalised in French. However, the formal concept **COLT** is a subconcept of the concept **FOAL**, which is lexicalised in French by the word *poulain*. The concept **COLT** has the definitional attributes {**horse, young, male**}, whereas the concept **FOAL** has only the definitional attributes {**horse, young**}. So defsur(**COLT**,**FOAL**) = {**male**}. The French lexicalisation of this attribute **male** is *mâle*. So the French word *poulain* is an incomplete translation of the English word *colt*. The part of the meaning that is missing is expressed in French by *mâle*, so taken together we get the complete translation *poulain mâle*, which is indeed what a colt is in French.

The process of lexical gap filling is exactly the same process as would take place when one tries to recover a monolingual definition from the system. To know the meaning of the English word *colt*, one has to look at the meaning it expresses; in this case mng(colt) = COLT. If we want to construe its definition in term of *genus proximum et differentiae specificae*, the genus proximum can be found by going to the first lexicalised superconcept of **COLT** (in this case **FOAL**), and then taking its $\text{wfs}_{\text{English}}(\text{int}(\text{FOAL}))$ = {foal}. We find the differentiae specificae by looking at the definitional surplus defsur(**COLT**,**FOAL**) = {**male**} and taking its lexicalisation (*male*). So the lexical definition for *colt*, which is *male foal*, is construed in exactly the same way as its translation in French. From this, we can deduce what the process of lexical gap filling renders: the lexical definition from the monolingual dictionary of the source language, translated bit by bit into the target language.

Although lexical gap filling renders the same result as generating a monolingual lexical definition, this is not necessarily the monolingual lexical definition that was put into the system in the first place. In fact, in our example it is not: the original definition for *colt* was *male young horse* and not *male foal* (see table 2.4)[12]. The reason for this is that the generation of a meaning description with the method described above is not a decisive process. It uses the notion of a lexicalised superconcept, but there can be various lexicalised superconcepts. In this case, also **HORSE** is a superconcept of **COLT**, with defsur(**COLT**,**HORSE**) = {**male, young**}. If we would have chosen this superconcept, its translation into French would have been *jeune cheval mâle*, and the English lexical definition *male young horse*, which would be the original LDOCE definition.

If we sharpen the process of lexical gap filling by saying that we should take the smallest superconcept, it would still not be decisive. Given the fact that the structure of FCA is not a hierarchy but a lattice, there does not have to be a unique smallest superconcept; not even a unique smallest lexicalised superconcept. If STALLION would not have had the attribute **adult**, then

[12]Notice that LDOCE hence does not use the genus *proximum*.

both STALLION and FOAL would have been immediate superconcepts of COLT, and there would have been no way of choosing between *male foal* and *young stallion* as a definition for *colt*. This non-uniqueness of result is not problematic: even though the processes of lexical gap filling and lexical definition generation do not yield a unique result, they do yield only correct results.

2.4 Formal Properties of FCA

The FCA system is not only simple, but also very powerful, and it has a lot of nice mathematical properties. These properties will be discussed in this section. Part of this logical analysis will be a calculation of the maximum number of formal concepts for a given context. The rest of this thesis will use, but not depend on, the properties described and proved in this section.

2.4.1 FCA and Lattices

One of the crucial properties of formal concepts is the fact that they form a complete lattice. That they do so has been proved by various authors, for instance in Davey & Priestley (1990), and Ganter & Wille (1996); let me reproduce the results here. First some standard properties of FCA (with $A_n \subseteq G$ and $B_n \subseteq M$), given in (2.10) - (2.14). All of these follow from the fact that by (2.14), we have a Galois connection, but can also easily be shown from the definitions of $A^{\uparrow}$ and $B^{\downarrow}$.

(2.10) $A \subseteq A'' \qquad B \subseteq B''$

(2.11) $A_1 \subseteq A_2 \Rightarrow A_2' \subseteq A_1' \qquad B_1 \subseteq B_2 \Rightarrow B_2' \subseteq B_1'$

(2.12) $A' = A''' \qquad B' = B'''$

(2.13) $(\bigcup_{t \in T} A_t)' = \bigcap_{t \in T} A_t' \qquad (\bigcup_{t \in T} B_t)' = \bigcap_{t \in T} B_t'$

(2.14) $A \subseteq B' \Leftrightarrow B \subseteq A' \Leftrightarrow A \times B \subseteq I$

With these, it is easy to show that $\langle \mathfrak{B}(G, M, I); \leq \rangle$ is a complete lattice since $\leq$ is an order, and the meet and join exist and are given by (2.15) and (2.16).

(2.15) $\bigwedge_{t \in T} \langle A_t, B_t \rangle = \langle \bigcap_{t \in T} A_t, (\bigcup_{t \in T} B_t)'') \rangle = \langle (\bigcup_{t \in T} B_t)', (\bigcup_{t \in T} B_t)'' \rangle$

(2.16) $\bigvee_{t \in T} \langle A_t, B_t \rangle = \langle (\bigcup_{t \in T} A_t)'', \bigcap_{t \in T} B_t) \rangle = \langle (\bigcup_{t \in T} A_t)'', (\bigcup_{t \in T} A_t)' \rangle$

A special interesting case given in (2.17); notice that this entails that any attribute that does not belong to any of the objects only appears at $\perp$.

(2.17) $\bigvee_{g \in A} \langle \{g\}'', \{g\}' \rangle = \langle A, B \rangle = \bigwedge_{m \in B} \langle \{m\}', \{m\}'' \rangle$

2.4.2 Smallest Common Concept

In section 2.1, page 29, I tried to show that $\langle A'', A' \rangle$ is the smallest concept for which all objects of A appear in its extent (the smallest common concept of A). Since we already know that $\langle A'', A' \rangle$ is a concept for any A (see (2.6)), we only need to prove that:

(2.18) $\forall A \in G.\ \forall \alpha \in \mathfrak{B}(G, M, I).\ A \subseteq \text{ext}(\alpha) \Rightarrow \alpha \geq \langle A'', A' \rangle$

This proof is straightforward: suppose $\alpha = \langle X, Y \rangle$ and $A \subseteq X$. Then $A'' \subseteq X''$ – by double application of (2.11) – and since $\langle X, Y \rangle$ is a concept, we know that $X = X''$. Hence $A'' \subseteq X$, and so by definition $\langle A'', A' \rangle \leq \langle X, Y \rangle$.

For single objects/attributes, the lowest common concept and greatest common subconcept are commonly written like this: $\tilde{\gamma} g = \langle \{g\}'', \{g\}' \rangle$ and $\tilde{\mu} m = \langle \{m\}', \{m\}'' \rangle$.

2.4.3 Maximal Filled Sub-Tables

In section 2.1 (page 23), the following claim was made: the set of formal concepts is identical to the set of maximal elements of the rectangles filled with crosses in the cross-table representation of the context. To prove this, firstly, define a set of filled rectangles $\mathfrak{F}$:

(2.19) $\mathfrak{F}(G, M, I) = \{\langle A, B \rangle \mid A \times B \subseteq I\}$

What we need to prove is that $\mathfrak{B} \subseteq \mathfrak{F}$, and more specifically, that $\mathfrak{B}$ is exactly the set of maximal elements of $\mathfrak{F}$:

(2.20) $\mathfrak{B}(G, M, I) = \{f_1 \in \mathfrak{F} \mid \forall f_2 \in \mathfrak{F}.\ f_1 \subseteq f_2 \Rightarrow f_1 = f_2\}$

Given the characterisation of $\mathfrak{F}$, we know that:

(2.21) $\langle A, B \rangle \in \mathfrak{F}(G, M, I) \Leftrightarrow \forall a \in A \forall b \in B.\ (a, b) \in I$

By the definition of $^{\downarrow}$ in (2.1), we have that:

(2.22) $A = B^{\downarrow} \Rightarrow \forall a \in A \forall b \in B.\ (a, b) \in I$

And hence that $\mathfrak{B} \subseteq \mathfrak{F}$. Any rectangle $\langle C, D \rangle$ that extends a concept $\langle A, B \rangle$ has to be identical to that concept. For it extends it iff $A \subseteq C$ and $B \subseteq D$; but in that case $\forall c \in C. \forall d \in D. (c, d) \in I$ and $\forall a \in A. \forall b \in B. (a, b) \in I$. So we can conclude that $D \subseteq B \subseteq D$, and $C \subseteq A \subseteq C$, which entails that $\langle C, D \rangle = \langle A, B \rangle$. QED.

2.4.4 Distributive and Atomic Lattices

In this subsection, some abstract properties of concept lattices will be discussed. The property that is most important for this thesis is that concept lattices are *atomic*: *"Ein vollständiger Verband, in dem jedes Element ein Supremum von Atomen ist, heißt atomisch."*[13] (Ganter & Wille, 1996 [7]). This property is important, since it is the underlying principle used for computing the formal concepts in JaLaBA (see section 2.5).

Atomic Lattices

Let **obpro** be the set of all sets of attributes that are projections by $^{\uparrow}$ of individual objects: $\{B{\subseteq}M \mid \exists a{\in}G.\ B = a^{\uparrow}\}$. This set **obpro** represents a set of concepts (the concepts $\langle \mathbf{obpro}', \mathbf{obbpro}\rangle$). And this set of concepts encloses the set of most specific concepts **atoms**, which are most specific in the following sense[14]:

(2.23) $\forall\alpha{\in}\mathfrak{B}(G,M,I).\ \alpha \in \mathbf{atoms} \Leftrightarrow \neg\exists\beta.\ \beta < \alpha \wedge \beta \neq \bot$

This is straightforwardly true: the concept represented by the set a' is $\langle a'', a'\rangle$, which we have proved to be the smallest concept having a in its intent (smallest common concept of a singleton). Since **atoms** contains such a smallest common concept for every element of G, the only concept that could be smaller than all of them is the concept containing none of the elements of G, or $\langle \emptyset, \emptyset'\rangle$ (which is only a concept (namely $\bot$) if $\emptyset'' = \emptyset$, otherwise put if there are no objects with all attributes).

The second claim is that the entire set of intents of all formal concepts (except possibly for $\bot$) is simply the set of all intersections of the elements of **atoms**. So let us define a set $\mathbf{atoms}^*$ of all inductive intersections of members of **atoms**: $x \in \mathbf{atoms}^*$ if $x \in \mathbf{atoms}$ or $x = y_1 \cap y_2$ and $y_i \in \mathbf{atoms}^*$, This set $x \in \mathbf{atoms}^*$ is closely related to $\mathfrak{B}(G,M,I)$:

(2.24) $\{B \mid \exists A.\langle A, B\rangle \in \mathfrak{B}(G,M,I)\} = \mathbf{atoms}^* \cup M$

The proof is trivial: on the one hand, every intent of a concept has to be the intersection of two others (which is a direct consequence of the definition of meet (2.16)), and on the other hand any intersection of two intents of concepts has to be an intent of a concept itself (which is also a direct consequence of (2.16)).

Since trivially we have that $\langle \mathbf{obpro}', \mathbf{obbpro}\rangle \in \mathfrak{B}(G,M,I)$, we can conclude that $\mathbf{obpro}^* \cup M = \mathbf{atoms}^* \cup M$. So we can conclude that concept

[13] A complete lattice, in which every element is a supremum of atoms, is called atomic. [my translation]

[14] Hence their name; in Representation Theory, atoms are the elements directly above $\bot$ (Davey & Priestley, 1990 [163]). The set **obpro** is not most specific in any sense: there can even be concepts not in **obpro** smaller than some concepts in **obpr**.

lattice are atomic, and that if we want to find the set of formal concepts, we only need to take the sets of properties of all the objects, and take all their intersections.

Distributivity

A *distributive lattice* has the following property:

$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c) \tag{2.25}$$

For concepts lattices to be distributive, this means that the following would have to hold:

$$\begin{aligned} &\langle A_1 \cap (A_2 \cup A_3)'', (B_1 \cup (B_2 \cap B_3))'' \rangle \\ &= \langle ((A_1 \cap A_2)'' \cup (A_1 \cap A_3))'', (B_1 \cup B_2)'' \cap (B_1 \cup B_3)'' \rangle \end{aligned} \tag{2.26}$$

That this does not hold is easily shown with a counterexample: take a context in which $G = (\{1, 2, 3\}$, $M = \{a, b, c\}$, and $I = \{(1, a), (2, b), (3, c)\})$. If we now fill in $A_1 = 1$, $A_2 = 2$, and $A_3 = 3$. Then $A_1 \cap (A_2 \cup A_3)'' = \{1\}$, while $((A_1 \cap A_2)'' \cup (A_1 \cap A_3))'' = \emptyset$. Ergo: concept lattices are not distributive.

2.4.5 Extending Contexts

If we have a context (G_0, M_0, I_0), and extend it to a context (G, M, I), with $\{G_0, M_0, I_0\} \subseteq \{G, M, I\}$, we do not want to lose the concepts we already had. So we need to prove the following:

$$\forall \langle A_0, B_0 \rangle \in \mathfrak{B}_0 . \exists \langle A, B \rangle \in \mathfrak{B}.\ A_0 \subseteq A \wedge B_0 \subseteq B \tag{2.27}$$

This is shown very easily. Since the new concept has to include the old one, it can only be the concept $\langle A_0'', A_0' \rangle$ (with the $'$ function from the larger context). This, as we have seen, will always be a concept. Furthermore, since $'$ yields the attributes that all objects in A_0 have, and $I_0 \subseteq I$, it logically follows that all the attributes in B_0 (which is A_0' with the function $'$ over the smaller context), will still be in A_0. So $B_0 \subseteq A_0'$. And since by definition (2.10) we know that $A_0'' \subseteq A_0$, we have proved (2.27).

There can be new concepts of course, and unless I is empty over both $A_0 \times B \backslash B_0$ and $A \backslash A_0 \times B$, there can be new concepts with existing objects (or attributes) in them.

2.4.6 Models and the Number of Concepts

As claimed earlier, there is no direct relation between the number of objects, the number of attributes, and the number of formal concepts related to a context. However, it would be interesting to know the following given a number of objects, and a number of attributes, what would be the minimum and maximum number of formal concepts, with a free choice of relation I? So given a context (G, M, I), what is the minimal and the maximal size of $\mathfrak{B}(G, M, I)$, for any I?

The minimum is trivial: if we take the universal relation, in which every object has every attribute, there will be precisely one formal concept: $\langle G, M\rangle$. And given the fact that $\langle X'', X'\rangle$ will be a concept for any X, there will always be at least one concept.

The maximum number of formal concepts is less easy to see. There are two cases that would have to be treated separately: either $|M| \geq |G|$, or $|G| \geq |M|$. But given the symmetry of the system, these cases run completely parallel, so let us assume that $|M| \geq |G|$.

Given the fact that any formal concept has a unique intent, and every intent is an arbitrary combination of attributes, it immediately follows that there is a total of $2^{|M|}$ possible intents. Therefore, no I could ever yield more than $2^{|M|}$ for a context (G, M, I) with $|M| = n$.

To show that this worst case of $2^{|M|}$ concepts also occurs, we need a context in which for every $B{\in}\wp(M)$, $\langle B', B\rangle$ is a concept. We get such a context if we take in injection $e{:}G \rightarrow M$, defined as $gIm \Leftrightarrow e(g) \neq m$; in other words, if we take a context in which every object relates to all attributes but one, and for every such almost complete set of attributes there is an objects that has all those attributes. With this, we have:

(2.28) **atoms** $= \{B \subseteq M \mid \exists g{\in}G.\ B = (e(g))^c\}$

Trivially, any $B \subseteq \wp(M)$ except M itself is an intersection of elements of **atoms**, so by (2.24), $\mathfrak{B}$ is exactly $\langle B', B\rangle$ for any $B \in \wp(M)$. The smallest such worst-case context is the bijection, where $|G| = |M|$.

This maximum number depends on the fact that $|M| \geq |G|$. Given the symmetry of the system, we can conclude that:

(2.29) $\max\limits_{I \in M \times G} |\mathfrak{B}(M, G, I)| = 2^{\min(|G|,|M|)}$

Multi-Valued Attributes

In the discussion above, we assumed that we could arbitrarily assign sets of attributes to objects. But more often than not, formal attributes are *not* independent: if something is blue, it will not at the same time be white or red. In these cases, we can see all colours as values for the same attribute. The

standard way of dealing with such multi-valued attributes is the following (Ganter & Wille, 1996): the context is extended to (G, M, W, I), where $I \subseteq G \times M \times W$, with $(g, m, w) \in I$ to be read as: object g has value w for attribute m, and $(g, m, v) \in I \wedge (g, m, w) \in I \Rightarrow v = w$. If W has n elements, it is called an n-valued context, and if $|M| = m$, there would be m^n possible intents, and hence concepts.

The disadvantage of this method is that there is one big, unordered set of values. However, in normal circumstances, every attribute will have its own set of values. An alternative is therefore to partition M into a set of sets, where every set contains the values of a specific attribute: **colour** can be *white*, *black*, or *red*, and **size** can be *big* or *small*. Notice that such a partition does not need to take semantics into account, but merely needs to obey mutual exclusiveness as in (2.30).

(2.30) $B \in \mathcal{P} \Rightarrow \forall g{\in}G.\forall b, c{\in}B.\ b \neq c \wedge gIb \rightarrow gIc$

With such a partitioned set of attributes, the extents of formal concepts still consist of sets of attributes, but not arbitrary ones: they have to be sets which contain for every $B \in \mathfrak{P}$ either one of its elements, or none of them. So every element of the partition gives $|B| + 1$ possibilities, and the total number of possible intents will be the product of all these possibilities. There is, however, exactly one extent in which all attributes, even from the same element of the partition, can coincide, namely $\perp$. So the maximal number of formal concepts will be:

(2.31) $\max_{I \in M \times G} |\mathfrak{B}(M, G, I)| = (\prod_{B \in \mathfrak{P}} |B| + 1) + 1$

As before, this does not have to be the actual number of formal concepts; take the toy model in figure 2.1, where we have two 4-valued attributes (**color** = *none, white, grey, black*; **form** = *none, square, round, circular*), so the maximum number of formal concepts would be $4^2 + 1 = 17$. That the actual number is lower (namely 13), is because this model accidentally has no non-white triangles, no white squares and no black or white triangles.

Notice also that where the basis of FCA is completely symmetrical in the sense that objects and attributes can switch places without much consequence, restrictions such as these are natural only for attributes, hence making it asymmetrical.

Additional Restrictions

Besides mutual exclusiveness, there are more properties that attributes can have, that limit the number of possible concepts. Notice that all these restrictions could in principle also be applied to objects, though that is less natural.

Coextensiveness When two attributes are coextensive, i.e. when $m_1'' = m_2''$, they can be counted as one.

Subordinateness Some attributes are subordinate: an attribute m_1 is subordinate to an attribute m_2, when it only appears within the context of that attribute, in other words when $\langle m_1', m_1'' \rangle \leq \langle m_2', m_2'' \rangle$. Such an attribute will contribute exactly one additional concept (m_2 is not also subordinate to m_1, for then they would be coextensive).

2.4.7 Partial Ordering on Attributes

The FCA framework assumes attributes to be completely independent. But sometimes, the attributes we use to characterise objects are just as interdependent as the concepts themselves. To give an example, if we have a model in which there are, amongst others, squares and rectangles, they would be the extent of different formal concepts, distinguishable by the fact that the squares are square and the rectangles are rectangular.

But being square is just a special case of being rectangular. It is being rectangular with the additional constraints that all the sides have to be of an equal length. So it is possible to divide the attribute of being square into two sub-attributes, namely being rectangular and having equally sized sides. But if we were to follow this line, we would have to further divide the attribute of being rectangular into having parallel sides that make corners of 90°, in order not to get into conflict with the attribute of being rhomboid (diamond-shaped; a square is also a special kind of rhomboid, where the sides make corners of 90°).

Although this line of action would 'solve' the problem, it would have us end up with attributes that are way too mathematical in nature to be useful in much everyday practice. We often just want to say that some objects are square, while others are merely rectangular, knowing that the first implies the second.

In that case, we have not simply a set of attributes, but a partially ordered set of attributes. This has consequences for the notion of being a sub-concept, since a concept can now also be a sub-concept of another if it has more specific, rather than simply more attributes. Adjusting the FCA framework for this new situation is rather trivial. In this definition, it is not strictly necessary, though useful, to first define the relation $\sqsubseteq$ of 'more restricting set of attributes', and redefine the relation $\leq$ with this new relation:

(2.32) $A_1 \sqsubseteq A_2 \Leftrightarrow \forall m \in A_1.\ \exists m' \in A_2.\ m \preccurlyeq m'$

(2.33) $\langle A_1, B_1 \rangle \leq \langle A_2, B_2 \rangle \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \sqsupseteq B_2$

In other words: a set of attribute is less restrictive than another, if every attribute of that is either itself a member of the more restrictive set, or there

is a stronger version of that attribute in the more restrictive set (2.33), and a more restrictive set of attributes yields a subconcept (2.32).

The new relation $\sqsubseteq$ in (2.32) is less strong than the subset relation $\subseteq$; it no longer defines a partial ordering, but merely a *quasi-ordering*. Although we still have transitivity ($A \sqsubseteq B \sqsubseteq C \Rightarrow A \sqsubseteq C$), and reflexivity ($A \sqsubseteq A$), we no longer have asymmetry ($A \sqsubseteq B \wedge B \sqsubseteq A \not\Rightarrow A = B$). We merely know that if for two sets we know for *both* that the first is more restrictive than the second, they are equivalent, but not necessarily identical:

(2.34) $A \sqsubseteq B \wedge B \sqsubseteq A \Leftrightarrow A \equiv B$

To take a concrete example; the set {**square**} is less restrictive that the set {**quadrangular, rectangular, square**}, for the simple reason that it is a subset. But the second is also less restrictive (not in the strong sense, but equally restrictive) than the first. The reason for this is that the additional attributes do not give additional information, since they are implied by the stronger attributes. But as a result, two non-identical sets can mutually be smaller in the relation $\sqsubseteq$ ($A \sqsubseteq B \not\Rightarrow A \not\sqsupseteq B$).

It is clear that this problem is created by the fact that there are 'useless' attributes in the set, and that if we would forbid those, the problem would go away. The best way to formally realise this, is not to remove all useless attributes, but instead to add them all. The downward closure of a set X (notated as $\downarrow X$) is exactly the set where we have added all the less informative elements, and we can redefine equivalence in term of downward closure:

(2.35) $\downarrow X = \{y \mid \exists x \in X.\ y \preccurlyeq x\}$

(2.36) $X \equiv Y \Leftrightarrow \downarrow X = \downarrow Y$

If we take the downward closure, we effectively get rid of the partial order on attributes: what we retain is that if $m_1 \sqsubseteq m_2$, that m_1 is subordinate to m_2. A special case of more restrictive attributes that we will encounter in the next chapter is one with disjunctive attributes; for instance, **flowing to the sea or another river** is less restrictive than **flowing to the sea**. This could be modelled logically:

(2.37) $\forall B_1, B_2 \in M.\ B_1 \sqsubseteq B_1 \vee B_2$

2.5 JaLaBA: an Online FCA Tool

Construing lattices from cross-tables is a completely automatic process, and can very well be done by a computer. As part of this thesis, I developed an online application that does this: the Java Lattice Building Application

(JaLaBA). JaLaBA an online tool that takes contexts as input, and renders their concept lattices. An illustration of the JaLaBA application is given in figure 2.8, where it has been used to create the lattice in figure 2.5. JaLaBA can be found on-line at the following URL:
`http://maarten.janssenweb.net/simullda`.
JaLaBA consists of two parts: a Perl script that asks for a context and builds an abstract representation of the corresponding concept lattice, and a Java applet that takes this abstract representation and turns it into an editable graphic representation of the corresponding Hasse diagram. These two parts will be discussed in turn here.

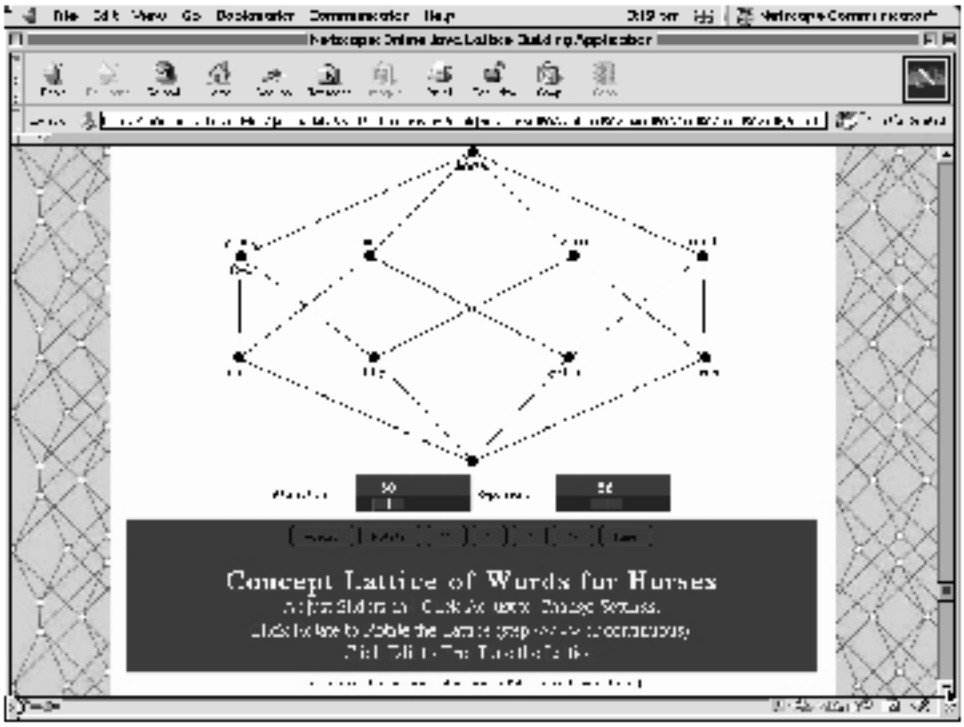


Figure 2.8: The Online Java Lattice Building Application

2.5.1 Construing Formal Concepts

The first requirement for building a concept lattice is having a definition of the underlying context. Since a context is little more than a cross-table, it can best be entered by means of a cross-table. But the number of formal objects and formal attributes is not predefined, therefore there is no way of telling the size of this cross-table. The method for dealing with this problem is to use HTML tables. By using an HTML based system, there is no need to worry about the interface itself, since the browser will take care of that:

it will display the table, process the input and pass it on to the rest of the program. And it has extendable windows that will adjust to the size of the table it has to display. The HTML table has to be generated on the basis of the desired number of objects and attributes, and this is done using a cgi-script. Since Perl is a very useful language for building cgi-scripts, the formal concept creation part of JaLaBA is written in Perl.

JaLaBA starts out by prompting for three arrays: an array for the list of formal objects, an array for the list of formal attributes, and a set of relations between them. It then puts the names of the objects and attributes aside, and treats the relation as relations between numbers. From this, it builds a comma-separated list for every object, of all the attributes that object has. These comma-separated lists form the basis for the generation of the lattice; this set of comma-separated lists is simply the set **atoms**, which constitutes the foundation of a lattice by means of equation (2.24): $\mathfrak{B} = \textbf{atoms}^* \cup M$.

Since **atoms** is a set, we can throw away all duplicate elements, by means of the following algorithm:

```
%found = "";
@atoms = grep { not $found{$_}++; $_ =~ s/,$//; }  @obj;
```

All elements of `@atoms` are then accepted as formal concepts, as well as all the intersections of elements of `@atoms`:

```
@fcatemp = @atoms;
for ($a = 0; $a < @fcatemp; $a++) {
  for ($b = $a+1; $b < @fcatemp; $b++) {
    %found = "";
    push(@double,join(",", grep { $found{$_}++; }
      split(/,/,$fcatemp[$a].",".$fcatemp[$b])));
  };
};
push (@fcatemp,@double);
push (@fcatemp,join",",(0..@attributes-1));
%found = "";  @fca = grep { not $found{$_}++; }  @fcatemp;
```

Although this will already be enough for most small contexts, it is not sufficient for layered concept lattices; and these can be rather small. In table 2.8, an example is given of a context for which it holds that not all the concepts of it will result from the procedure described above (the largest set of formal concepts for 3 objects and 3 attributes, resulting in a cube): the top-concept $\top$ will not found.
The reason for this is that the procedure only yield intersections of pairs of set. But these intersections only form the first superconcepts of **atoms**; in a multi-layered lattice, there will be further superconcepts of the intersection. These have to be found by further taking the intersections between the new concepts generated in the previous step with the existing with the existing

	a	b	c
1	×	×	
2		×	×
3	×		×

Table 2.8: Small Layered Context

elements of `@fcatamp`. This because the newly added sets of attributes in the last step can in principle lead to more FC's, but only if two conditions are met: firstly, the new set has to be non-empty, and secondly, the set has to be really new, i.e. not yet present in `@fca`. The solution to this is to add new elements meeting these conditions on the fly to the set `@fcatamp`. So the complete algorithm for finding all the formal concepts is the following:

```
%found = "";
@fca = grep { not $found{$_}++; $_ =~ s/,$//; }  @obj;
@fcatemp = @fca;
for ($a = 0; $a < @fcatemp; $a++) {
  $fcadone{$fcatemp[$a]}++;
  for ($b = $a+1; $b < @fcatemp; $b++) {
    %found = "";
    $double = join(",", grep { $found{$_}++; }
      split(/,/,$fcatemp[$a].",".$fcatemp[$b]));
    if ( $temp ne "" && not $fcadone{$double} ) {
      push(@fcatemp,$double); $fcadone{$double}++;
    };
    push(@double,$double);
  };
};
push (@fcatemp,@double);
push (@fcatemp,join",",(0..@attributes-1));
%found = "";  @fca = grep { not $found{$_}++; }  @fcatemp;
```

Unfortunately, there is no formal proof that this larger procedure does indeed yield all formal concepts: the Perl language has no formal semantics. Therefore it is impossible to prove that this procedure actually yields the complete set of formal concepts.

After the formal concepts have been found, the partial order on these concepts has to be established[15]. Given the fact that all concepts are simply comma-separated lists of numbers, this is easy: for two lists a and b in `@fca`, a is a superconcept of b if all the elements of a are also elements of b.

[15]In a more efficient design of this algorithm, such as proposed by Kamps (1997), this would not be necessary: the partial order was already clear in the previous step. All elements of **atoms** are larger than $\perp$, and all intersections are larger than both their composing parts.

If we thus list all superconcepts for every element of `@fca`, this will give an abstract representation of the lattice, consisting of a list of pairs, where the first element is the name of the node in question, and the second element is a list of all its superconcepts. The abstract representation of the lattice in figure 2.8 is given in figure 2.9.

This abstract representation of the lattice is formatted in such a way that it can serve as the input for the second part of the lattice building: the Java applet drawing the actual lattice.

2.5.2 Drawing Lattices

The second phase of drawing a concept lattice is creating a Hasse diagram on the basis of the abstract representation of the lattice produced by the Perl-scripts. Drawing a Hasse diagram has an inherent problem: by the way they are defined, Hasse diagrams only give the vertical order of the nodes. There is no specification about their horizontal configuration. Drawing a diagram that simply obeys the specifications is easy. But drawing a 'good' diagram is less so:

> Finding a 'good' diagram is not an entirely mathematical problem since the idea of a 'good' diagram is, at least in part, æsthetical. On the other hand, a poorly chosen diagram of even a small, well known lattice can render it unrecognizable ... An example of a diagram which makes it hard to recognize the [8 element Boolean algebra = Cube] lattice is given in figure 2.10. (Freese *et al.*, 1991 [251])

```
(   (0   ())
    (1   (0 6 7 ))
    (2   (0 6 8 ))
    (3   (0 ))
    (4   (0 3 7 ))
    (5   (0 3 8 ))
    (6   (0 ))
    (7   (0 ))
    (8   (0 ))
    (9   (0 1 2 3 4 5 6 7 8 ))
)
```

Figure 2.9: Abstract Representation of Lattice in Figure 2.5

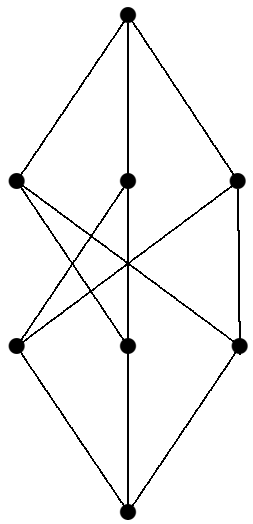


Figure 2.10: Concept Lattice of Cube

To generate diagrams, JaLaBA uses a Java applet that takes HTML input, and then produces the desired result. This Java applet is based upon the *Lattice Draw* applet, written by Ralph Freese. JaLaBA is a slight adaptation of Lattice Draw, with some additional features. More importantly, JaLaBA adds the Formal Concept Analysis labels for formal objects and formal attributes above and below the appropriate nodes.

Lattice Draw organises the elements in the lattice by building a 3D model of it. This is done in three steps:

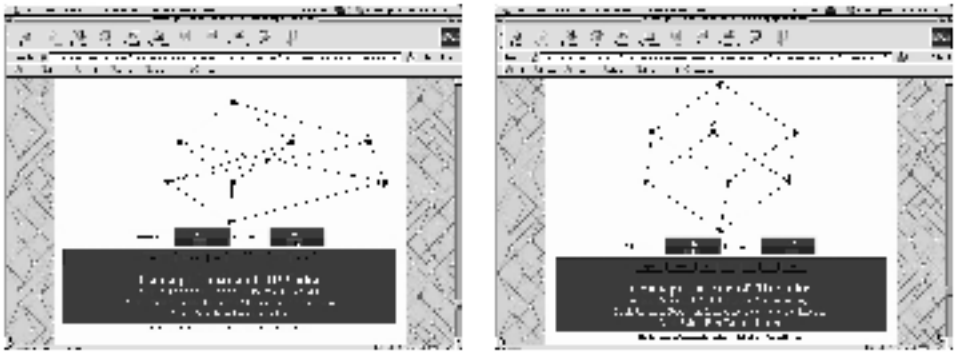
Height Function For a lattice **L** with k as its longest chain, $\text{ht}(a) = r - s + k$, where r is the longest chain from $\perp$ to a, and s the longest chain from a to $\top$

Position in 3D Space The set H of element with height h are placed on the plane $y = h$, initially in a circle, with $x^2 + y^2 = |H|$

Forces All similar points (i.e. points that are connected) attract each other proportional to their distance, with forces only operating in the $x - y$ plane; and all dissimilar point equally repell each other in the same fashion. The point is then moved by these forces over a distance proportional to $\sqrt{n}$, where n is the size of the lattice. This process is repeated untill an equilibrium is reached.

These three steps yield a 3-dimensional arrangement of all the nodes in the lattice. This 3-D object can of course be projected onto a 2-D space, giving a Hasse diagram of the lattice. But there is no guarantee that this projection

will be a 'nice' lattice in the intended sense. However, the object can be rotated and projected under any angle. This gives a large range of accessible representations for the lattice, amongst which a 'nice' lattice is bound to appear. Figure 2.11 gives an illustration of how an ugly representation of a cube can be changed to a nice representation. Furthermore, the rotating object itself does not merely look good, but has an additional advantage: for more complicated lattices (such as the 4-dimensional hypercube), it is hard to understand the structure of the lattice under any angle, and the object can best be viewed as a rotating 3-D object.



Before After

Figure 2.11: Adaptation of a Lattice

There are of course more direct ways of assuring a 'nice' representation of a lattice. For instance, Ganter has argued that for a nice representation, two things should be minimized: the number of crossing connections, and the number of different angles in the diagram. This last point maximizes the number of parallel lines, which is one of the main criteria for 'niceness'. The JaLaBA applet does not include such rules, since it builds its lattice in 3D, while such rules always apply to 2D images. Furthermore, the rotatability of the Lattice Draw set-up usually yields a 'nice' image, and otherwise the nodes can be moved manually.

2.6 Conclusion to Chapter 2

In this chapter, I have introduced the basic principles of Formal Concept Analysis, both formally and intuitively. In both cases, formal concepts are pairs of objects and their attributes, that belong together because the objects share all the attributes (and vice versa). But intuitively, we can also say that formal concepts are maximal areas of filled squares in a cross-table.

The logical system of Formal Concept Analysis can be applied to dictionaries in the following way: take the (disambiguated) words as formal

objects, and their dictionary definitions as formal attributes. This application of FCA to the content of dictionaries is the foundation of SIM*u*LLDA.

With this set-up translations for words can be found: find the interlingual meaning connected to the citation-form of the desired word-form, and see if there is also a citation-form related to it in the target language. If so, this will be the translation of the word. More interestingly, a translation can also be construed if no lexicalisation of the interlingual meaning exists in the target language (i.e. if there is a lexical gap): find the first superconcept of the smallest common concept of the interlingual meaning for which a translation is required. Then, find the definitional surplus of the smallest common concept w.r.t. this superconcept. The lexicalisation of the superconcept, together with the lexicalisation of the definitional surplus will be the desired translation; it will be an explanatory equivalent and not a translational equivalent. The monolingual definitions of words can be found in the same way, by taking source and target language to be the same.

In the formal-properties section, I have shown that in a worst case scenario, the number of formal concepts grows exponentially with the number of attributes (2^n). But in lexicographic contexts, attributes normally have features that restrict the number of formal concepts: attributes can coincide on every occasion; appear only in the context of some concepts, or form sets of mutually exclusive attributes, i.e. be part of a multi-valued attribute, like colours are.

Much of the actual application of SIM*u*LLDA depends on the precise way in which the basic elements of the system are interpreted: words, languages, interlingual meanings and definitional attributes. Therefore, the next chapter will be dedicated to a thorough analysis of the appropriate interpretation of these elements.

Chapter 3

SIM*u*LLDA Elements

As explained in the previous chapter, the semantics of the FCA system is entirely dependent on the interpretation of the basic elements. Therefore, in order to have a precise meaning for the system, a precise description of these basic elements is required. Of course, fine-tuning the interpretation will not affect the system as such: if we fill the SIM*u*LLDA system with lexicographic data, the lexical gap filling procedure described in the previous chapter will generate bilingual definitions. And these bilingual definitions will not change by careful consideration of the nature of the basic elements that played a role in their conception.

But there are two reasons for a detailed analysis of the basic elements. Firstly, word-meaning is a very difficult field, and there are many problematic cases. To reach a proper analysis of these problematic cases, it is necessary to have a precise perception of what every analysis implies. And secondly, there are many aspects and elements of dictionaries that are not accounted for by the basic set-up explained in the previous chapter. Given the complexity of semantics, I believe it is only possible to deal with additional elements properly and at the appropriate level. This chapter will be dedicated to the analysis of the nature of the basic elements of the theory: words, interlingual (word) meanings, languages, and definitional attributes.

3.1 Words and Word-Forms

The first component of SIM*u*LLDA to look at in more detail is the set of 'words'. Words in the SIM*u*LLDA set-up are language-dependent elements, grouped into languages, and linked to one or more interlingual meanings in the concept lattice. But for a good interpretation we need a more formal and complete definition of what 'words' in SIM*u*LLDA are. To answer this question, I will give an overview of what types and aspects of words we can distinguish. This is a matter that has been discussed at length in

various fields such as linguistics, lexicography, philosophy, and psychology. The discussion in this section will follow the standard literature on the subject, but deviate from it where need arises. As part of this general specification of words, I will turn to the question which of these aspects of words should be part of the SIMuLLDA set-up, and where and how they should be modelled. This section is in part a modification of (Janssen & Visser, to appear).

3.1.1 Word-Form and Lexeme

Because dictionaries are all about words, and this thesis deals with the content of dictionaries, it should not be surprising that the word 'word' occurs often. It is, however, much too vague a notion to be used in a formal context. The word 'word' does not have a clear, single meaning: it is used loosely for a written word, a spoken word, a word-meaning, etc. Sometimes, we even refer to entire sentences as words (*"In one word: it was absolutely wonderful."*). I will, therefore, avoid in the following the use of the word 'word' as a technical term and use it only informally. Instead, I will use intuitively less clear, though formally much better defined notions. This section will introduce some common terminology, most of which is drawn from the standard work on semantics by Lyons (1977; 1995), with the addition of a few more technical terms. On top of the terminology, some notational conventions will be introduced, that will be used throughout this thesis, to keep the different kinds of 'words' apart. When just the abstract <u>word</u>-as-such is meant, it will be <u>underlined</u>.

The definition of a word that usually springs to mind first is that of a sequence of letters, also called a *string* or an *orthographic word*. `Strings` will be typeset in `courier`. The identity of strings is straightforward: two strings are identical iff they consist of the same sequence of letters[1]. Strings come in *types* and *tokens*: every use of the word 'the' on this page (for instance) is in some way a different word (token string), although they are also instances of the same word (type string).

A word can also be a spoken unit, or a *phonological word*. The identity of phonological words is much less straightforward than that of strings: where words (nowadays) have a prescribed spelling, a word is hardly ever pronounced exactly the same way twice, and pronunciation can significantly vary across dialects. A phonological word is commonly identified with a certain 'prototypical' or commonly accepted pronunciation, often represented in the International Phonetic Alphabet (IPA) in dictionaries[2]. Phonological words will be typeset in IPA between /slæsʃəs/. Also phonological words come in types and tokens.

[1]This, of course, pushes part of the problem down to the identity of letters.

[2]The pronunciation of IPA characters is given in appendix B.2. IPA also can be used for tone languages like Chinese.

The English word medieval has two different ways of spelling: either as `medieval`, or as `mediaeval`. In the same fashion, there are also two ways of pronouncing it: either as /medi'iːvl/ or as /miːdi'iːvl/. Given the definitions of orthographic and phonological words, this means that there are both two distinct (type) orthographic words, and two distinct phonological words. However, in a more interesting way, we say that they are all related to a single 'word'.

An abstract model of some salient aspects of a word that cuts across spelling and pronunciation is provided by the *word-form*. Word-forms will be typeset in sans-serif. As exemplified above, a word-form can have spelling and pronunciation variations. But conversely, different word-forms can also have the *same* spelling, for instance the Dutch word-forms band_1 (band) and band_2 (tyre) are pronounced respectively as /bɛnd/ and /bɑnt/, but both spelled as `band`. The distinguishing feature between these two word-forms is thus the pronunciation. Such distinct word-forms that have the same spelling are called *homographs*, whereas word-forms that are pronounced in the same way (like lesson and lessen) are called *homophones*[3].

But two word-forms can even be homophones *and* homographs at the same time, while still being distinct. The word-class distinguishes the noun hammer from the (related) verb spelled and pronounced the same way. In a gender-sensitive language like Dutch[4], you find differences between a neutral word bal (ball; party) and a male word bal (ball; sphere). Though the notion of a word-form is often treated as a well-defined concept, there does not seem to be a clear definition of its identity criteria. Here is an attempt at a definition.

Word-Form A *pre-word-form* is given by a spelling-cum-syllabification and a pronunciation.
A *word-form* is given by a number of pre-word-forms plus a word-class plus a gender (if applicable).

We can model the notion of word-form as a tuple of a set (viewed disjunctively) of pre-word-forms, a wordclass and, possibly, a gender. With the

[3]These notions can also be used cross-linguistically: the Dutch word *mais* (corn) is a homograph of the French *mais* (but), and a homophone of the English *mice*.

[4]Dutch does not heavily use gender though.

word medieval (as a noun) we may associate the following word-form:

$$\begin{array}{ll} \langle \ \{ & \langle \mathtt{me \cdot di \cdot e \cdot val},\ \text{/medi'iːvl/} \rangle, \\ & \langle \mathtt{me \cdot di \cdot e \cdot val},\ \text{/miːdi'iːvl/} \rangle, \\ & \langle \mathtt{me \cdot di \cdot ae \cdot val},\ \text{/medi'iːvl/} \rangle, \\ & \langle \mathtt{me \cdot di \cdot ae \cdot val},\ \text{/miːdi'iːvl/} \rangle \ \ \}, \\ & \text{count noun}, \\ & \text{male} \ \rangle \end{array}$$

Since set-theoretical notation does not make for pleasant reading, we replace it by a more convenient box format. The above representation thus becomes:

me · di · e · val	/medi'iːvl/
me · di · e · val	/miːdi'iːvl/
me · di · ae · val	/medi'iːvl/
me · di · ae · val	/miːdi'iːvl/
count noun	
male	

The representation of medieval as an adjective will be:

me · di · e · val	/medi'iːvl/
me · di · e · val	/miːdi'iːvl/
me · di · ae · val	/medi'iːvl/
me · di · ae · val	/miːdi'iːvl/
adjective	

The reason to have pairs of spelling and pronunciation is to allow for the possibility of spelling and pronunciation to be linked here: it is explicitly modelled that a certain spelling (say, the American spelling) can only be pronounced in certain of the possible ways (the American ones). If we ignore this possibility, we can represent a word form by a tuple of a set of spellings-cum-syllabification, a set of pronunciations, the word class and, if necessary, the genus. E.g. our word-form for medieval$_{\text{noun}}$ would become:

me · di · e · val me · di · ae · val
/medi'iːvl/ /miːdi'iːvl/
count noun
male

Dictionaries, given their textual presentation, have to rely on this simpler model; in many dictionaries, spelling variation is represented by having various comma separated possibilities, and pronunciation variation by having verticle-bar separated versions, as in the LDOCE entry for medieval:

medieval, mediaeval /ˌmediˈiːfəl||ˌmiː-/ *adj* **1** of the period in history between about 1100 and 1500 (the MIDDLE AGES) **2** *infml derog* very old or old-fashioned

Arguably our present representation is not quite correct. E.g. the Dutch word idee can be both male and neutral, as one of the many variations of words, as exemplified in table 3.1. This is easily repaired by allowing more than one gender.

band	bɛnd	band
band	bɑnt	tyre
sponsoren	spon·so·ren	sponsors
sponsoren	spons·o·ren	*sponge ears (mock word)
negeren	ˈne·ge·ren	to torment
negeren	ne·ˈge·ren	to ignore
blik[5]	m	glance
blik	n	can
leren	adj	leather
leren	verb	to learn

Table 3.1: Types of Homographs in Dutch

Note that the word-form is not identical to the word: it is a standardised way to represent certain crucial features of a word.

Lexemes

We can build an even more abstract notion. We say that *is* and *are* are different word-forms, but still in some way the same: they are morphologic expansions or *inflections* of the same *word-expression*. Thus, the rows in table 3.2 contain different word-forms of the same word-expression.

For English, every word-expression has only a few word-forms. However, if you consider a much more inflectional language like Czech, word-expressions can have lots of inflections, an example of which is given in table 3.3.

book	books	book's	books'	
mouse	mice	mouse's	mice's	
great	greater	greatest	greatly	
bad	worse	worst	badly	
look	looks	looked	looked	looking
go	goes	went	gone	going

Table 3.2: Inflection in English

	Sing	*Plur*	*P.M.Sing.*	*G.F.Sing.*	*F.Plur*
Nom	bratr	bratři	bratrův	bratrova	bratrovy
Acc	bratra	bratry	bratrova	bratrovu	bratrovy
Gen	bratra	bratrů	bratrova	bratrovy	bratrových
Dat	bratru	bratrum	bratrovu	bratrově	bratrovým
DL	bratrovi	bratrech			bratrových
Ind	bratrem	bratry	bratrovým	bratrovou	bratrovými
Voc	bratře				

Table 3.3: Inflection of bratr (brother) in Czech

By convention, every word-expression has a *citation-form*. Which of the word-forms counts as the citation-form depends on the language and the word-class. The citation-form for verbs in English is the infinitive, for adjectives in Czech it is the singular male nominative. These citation forms are the *headwords* of definitions in dictionaries. However, headwords of dictionary definition are more generally citation-forms of *lexemes* rather than word-expressions. *Lexemes* will be typeset in *slanted* fonts. Lexemes are strictly speaking not words but *phrases*, since there are also *multi-word lexemes*: semantically inseparable units consisting of more than one word, such as *pass muster* and *food poisoning*. Also idiomatic constructions such as *red herring* (in its non-literal meaning) are considered multi-word lexemes.

The lexemes can be said to be represented by their citation- or entry-forms: "*The lemma, when used as an entry-form, conventionally represents all the inflected forms of the unit,* umbrella *for* umbrella *and* umbrellas, *take for* take, takes, taking, taken, *and* took, *or go for* go, goes, going, gone, *and* went*: inflected forms are normally all treated together in the same entry, under the same entry-form.*" (Béjoint, 1994 [192]).

Individuation Problems

The standard definitions of lexemes, strings, phonological words, and word-expressions work smoothly for western-European languages. But there are languages that have additional complications. For instance, the notion of

an orthographic word is less straightforward in Japanese. First of all, where English has only 26 characters to compose words, Japanese has some 5000 kanji characters. But this is more a practical than a theoretical problem. But on top of this, Japanese has also three independent writing systems: kanji, hiragana and katakana. Kanji is the Japanese version of Chinese characters, whereas hiragana and katakana are both syllable-based writing systems. The same word can be written regularly either in Kanji or in hiragana. An example of this is given in figure 3.1, where the kanji and hiragana version of the Japanese word for *horse* (uma) are given.

馬 うま /ˈuma/ **horse**

Figure 3.1: Kanji and Hiragana versions of /ˈuma/

The question in this example is whether these two derivable ways of writing the same word should be considered different orthographic words; the hiragana version is in some sense a transcription of the first, much like Chinese can be transcribed into Pinyin notation, or English words into IPA. A similar problem arises with Serbo-Croatian, which can be either written in Cyrillic, or in Roman characters.

Although not entirely satisfactory, strings will still be viewed as sequences of characters in these cases, and kanji and hiragana, or Roman and Cyrillic simply lead to different characters. So even regular spelling variation still has to be treated individually for every word, although the variants themselves can of course be generated automatically if predictable.

Lexemes in Simullda

The purpose of the SIM*u*LLDA system is to fruitfully model the content of dictionaries. And since the lexical entries in dictionaries concern headwords, that is logically also what the language-modules in SIM*u*LLDA will consist of: citation-forms, which are word-forms representing lexemes.

Notice that the same word-form can represent different lexemes. For instance, the French word juste (adj.) can either be expanded as juste-justement-justesse, or as juste-justement-injuste-justice-injustice (Messelaar, 1990 [39]). Since there are no formal differences between these two word-forms juste, they are the same word-form representing completely different lexemes.

Which word-forms are represented by the citation-form is a question that we will get back to in section 5.1.1. For the moment, it suffices to say that lexemes contain word-forms with the same meaning, as formulated by

Mel'čuk:

> *Grosso modo*, un lexème est une unité du plan paradigmatique: l'ensemble maximum des mots-formes qui ont tous une même signification (le plus souvent) lexicale et, le cas échéant, une même signification dérivationnelle, mais qui manifestent des significations flexionnelles bien différentes et qui, par cela même, s'opposent entre eux, en s'excluant mutuellement dans une position donnée dans le texte[6].
> (Mel'čuk, 1993 [337])

Notice that word-forms are not identified with the Saussurian notion of a *signe*; according to de Saussure, a signe encompasses both a *signifiant* (roughly the word-form) and a *signifié*: its meaning. So the elements that de Saussure takes as primitives are word-senses and not word-forms. As will be discussed in section 3.3.1, word-senses are difficult for various reasons; however, they are also necessary in many senses. So word-senses will be present in SIMuLLDA, as will be explained in the next chapter. Word-senses are already implicitly present in the relation between the citation-forms and the interlingual meanings.

3.1.2 Morphemes

Although words have a rather intuitive status, many linguists have argued that they have no scientific value, because space-separation is not a relevant criterion. On the one hand because larger units, like multi-word lexemes and noun-phrases, behave almost like words. But more significantly, because words are not atomic units as grammatical and semantic entities. This is nicely illustrated by looking at polysynthetic languages. For instance, the Inuktitut expression *Qasuiirsarvigssarsingitluinarnarpuq* consists of just one word, but has the following internal structure[7]:

(3.1)

qasu-irr	-sar	-vig	-ssar	-si
tired-not	*cause*	*place-for*	*fit*	*find*
-ngit	-luinar	-nar	-puq	
not	*completely*	*somebody*	*3rd pers. sing.*	

Less exotically, Dutch and German compound nouns are written as a single word: koeienmelk is compositionally build up of melk (cow) and koeien (of a cow). And the Dutch word optillen (lift) is split into two parts and spread

[6] *Grosso modo*, a lexeme is an element at the paradigmatic level: the maximal set of word-forms that have a same (mostly) lexical meaning and, when need arises, a same derivational meaning, but that manifest very different inflectional meanings and that, by themselves, oppose one another, by mutual exclusion in the position given by the text.

[7] Example taken from Hendriks *et al.* (1997 [183]).

across the sentence when used in non-infinitive form: *ik til iets op* (I lift something).

In these examples, the atomic units are not words, but parts of words. Therefore, most linguists have discarded the notion of a word as a theoretically uninteresting entity, and focus in theory only on the in many ways more relevant notion of morphemes.[8] Morphemes and words can coincide, and in languages without affixes like Chinese, they even always coincide.

The fact that in the previous section words in SIM*u*LLDA were identified with lexemes (consisting of word-forms) seems to be at odds with the rejection of words in favour of morphemes. But although the composing elements of lexemes are called word-forms, they are not 'words' in the sense of space-separated units. The possibility of multi-word lexemes was already mentioned, in which case units larger than 'words' are treated like lexemes. In the same way, morphemes that are smaller-than-word units can be lexemes. This is not uncommon in dictionary practice, as described by Zgusta:

> The fact that we regard the words (*qua* lexical units) and multiword lexical units as the basis of the lexicographer's selection does not, however, mean that we should leave units smaller than the word out of consideration. In those languages where the boundary of the word is not sufficiently clear, the lexicographer will meet morphemes about which he cannot easily and unequivocally decide whether they are words of their own or not; in the majority of cases he will be well advised to allow them their own entries as if they were independent words, eventually with some further special indications and specifications. But sometimes the lexicographer also indicates as an entry a mere morpheme even if there is no uncertainty about the word boundary and when it is clear that it is only a morpheme. This is the case, for example, of highly productive prefixes or compositional elements. For instance, a prefix like anti- or pseudo- is so highly productive in many European languages, that it is impossible to indicate all the instances where it occurs; even if they were listed, new creations, and many of them occasional, could be expected to arise at any moment. Therefore, it is fully legitimate to indicate single morphemes, i.e. in this case the isolated prefix, describe its meaning, and add some of the more important and stabilized words in which it occurs. (Zgusta, 1971 [241])

So morphemes are not discarded, but morphemes, words, and multi-word lexemes are all treated equally, and all referred to as word-forms. Space-separation is not a criterion for word-hood.

[8] A very thorough analysis of the lack of (grammatical) value of words has been given by Marit Julien in her thesis (Julien, 2000).

3.1.3 History, Etymology

In the definition of a word-form, many of the aspects of words were taken as part of their identity criterion. But one of the aspects that commonly also distinguishes word-forms in dictionaries was not taken as such: the history of the word.

Many, if not all, words change their form and meaning over time. Still, we want to say that the word outrageous as used by Shakespeare in *Hamlet*[9] is the same word as the word outrageous we use today. This is related to a well-known ontological problem: the human body replenishes all its atoms about every 10 years. So no single part of it remains the same, yet still we want to say it is the same body. The same can happen to words: a well-known example is the word nice, whose history can be found in the OED. It is derived from the word *nescius* (ignorant), and used to be a negative word: in the 14th and 15th century it meant *foolish* or *stupid* (*they seiden he was a fool…and that they sien neuere so nise a man*, ca. 1450). By 1529, it got to mean *wanton*, or *loose-mannered*. It then shifted to *tender*, *reserved*, and *requiring great precision*, which lead in the 18th to *dainty* or *appetising*, esp. of a cup of tea (*we sent her up three or four plates of the nicest things that were at table*, 1766). The meaning it now has (agreeable) is incomparable with its original meaning, and its spelling changed from `nise` to `nice`. So no aspect of it remained, yet still we want to call it the same word.

Although hard to formalise exactly, the identity of the word seems to follow the idea behind the identity of the human body: as long as it can be continuously traced back to its root, we keep considering 'it' the same word.

One problem with this method of identification of words can be illustrated by loanwords: the English word apartheid can be continuously traced back to the Dutch word apartheid; and we could easily imagine considering these the same word. But now take this one step further: the old Dutch word mannekijn (little man) was used for the dolls to show new fashion. This word was adopted by French as a loanword and written in a more pronounceable French way as `mannequin`. If history would determine the identity of words, than mannekijn and mannequin would be the same word. But the dolls over time were replaced by women showing fashion and in its French spelling adopted as a loanword in Dutch, as described in LDOCE:

> **mannequin** [ledenpop](19e eeuws) < **fr.** *mannequin* < **middelnl.** *mannekijn* [mannetje, pop]; tot in de 19e eeuw werd damesmode bekend gemaakt door het sturen van poppen.

The same hold for the word bolwerk (fortification), which was adopted into

[9] To be or not to be, that is the question, whether it's nobler in the mind to suffer the slings and arrows of outrageous fortune…

French, changed it meaning to the broad road around the fortification, and readopted back into Dutch. But clearly the Dutch words bolwerk and boulevard are different.

A dictionary is always a temporal 'projection' of the language, in which the current meaning and form of the words is described. History, in a way, plays only a marginal role in them.

3.2 The Language

Even more than the notion 'word', the notion 'language' seems to be an intuitively clear notion. In the previous section, it was established that languages in the SIMuLLDA set-up consist of citation-forms of lexemes. However, there still remains the question *which* lexemes belong to the language modules. Or: what is a language? There are two separate questions to be answered in that respect, the first is: when do we call something a language instead of a dialect; Korean and Spanish are two different languages, so much is clear. But how about British English and American English? And is Old English the same language as modern English and if not, when did the one become the other?

The second question is: which words do belong to the language? How frequent, wide-spread, current, and established does a word have to be in order to count as part of the language? In terms of the SIMuLLDA set-up the two questions that this section will address are: what languages modules do we need, and which words go into them? In this section I will show that neither of these questions has a definite answer. Consequently, strictly taken a language is little more than an arbitrary list of lexemes.

3.2.1 Language, Dialect and Idiolect

Every language has dialectic variance: the pronunciation of the words, the meanings of the words, the words themselves, their spelling and even grammar are not homogeneous throughout the population of a language. Many languages (or countries) have a commonly accepted standard dialect, such as BBC English or ABN (common civilised Dutch), but regional, social or professional groups can deviate from this standard on any of the aforementioned aspects.

Dictionaries, recognising this variation, usually mainly describe the commonly accepted dialect, and only marginally describe dialectic variation. Where dialectic words or meanings are given, they are always indicated by a label. For instance, LDOCE uses labels like AmE, BrE, CanE, SAfrE to mark words specific for respectively American, British, Canadian and South African English (there are 11 distinguished dialects in LDOCE). This dictionary practice implicitly assumes that it is possible to distinguish di-

alects from languages: the Dutch DvD does contain words from Flemish, since Flemish is considered a dialect of Dutch. However, it does not incorporate words from Afrikaans, since Afrikaans is not (or no longer) considered a form of Dutch, but rather a separate language. In this section, I will argue that no such hard division is possible.

The standard definition for distinguishing languages from dialects is the following: a language is a collection of mutually intelligible dialects, so when two dialects are not mutually intelligible, they belong to different languages. But as any introductory textbook will explain, this definition does not work (even though it is often used for language classification in practice). There are a number of reasons why it does not work, amongst which the following two: mutual intelligibility is not a transitive relation, and it also is a graded relation. I will show why these are problems, and illustrate them with some well-known examples (see for instance Wardhaugh (1986) or Gleason (1955)).

The cited standard definition defines a language as what logicians call an *equivalence class* on dialects, with intelligibility as its equivalence relation. An equivalence relation is a reflexive (every dialect is intelligible to itself), symmetric (if X is intelligible to Y, then Y is also intelligible to X), and transitive relation (if X is intelligible to Y, and Y to Z, then X is also intelligible to Y). But the relation of (mutual) intelligibility does not have all these properties: although intelligibility is reflexive, it is not entirely symmetric (Danes understand Norwegians better than vice versa), and definitely not transitive. A good example is that the dialects on the border between Germany and the Netherlands are mutually intelligible across the border, and on both sides of the border they are mutually intelligible with standard German and Dutch respectively. But German and Dutch are not mutually intelligible.

Also, the definition expects mutual intelligibility to be an all-or-nothing relation, but it is not. Although a monolingual English speaker will not understand a word of Vietnamese, closer dialects vary from slightly comprehensible to (almost) fully intelligible. Italians recognise a lot of the words in Spanish, but cannot fully understand it, while Polish and Russian people can almost understand each other if they talk slowly. Swedish and Norwegian are even better intelligible, and even often better intelligible than dialects of the same language are: for many Dutch people (or at least for me), Norwegian is about as intelligible as the Dutch dialect spoken around Ieper in Belgium. Hardly any dialects are fully mutually understandable: although British and American English are usually very much mutually intelligible, in crucial circumstancesa 'translation' between the two can be necessary. This is nicely illustrated by the fact that during WWII, when the English and Americans were fighting side by side, both the English and the American military issued dictionaries to help bridge the language gap. This mostly to overcome the lexical differences as illustrated in table 3.4.

British English	American English
colour	color
luggage	baggage
toilet	rest room
pavement	sidewalk
lift	elevator
cock	rooster
trade union	labor union
pissed	drunk
annoyed	pissed

Table 3.4: British vs. American English

This problem of mutual intelligibility gets even more complicated if you consider Chinese: Mandarin and Cantonese are not very much mutually understandable (or even not at all) when spoken. However, they have a shared writing system, which makes them perfectly understandable in written form[10].

The fact that the standard definition does not work does not by itself mean that no proper definition can be given. But given the complexity of the problem, and the many borderline cases involved, it will be very hard to give a clear and satisfactory definition. In the actual discrimination of language from dialect, often politics plays an important role. Until about 1920, Afrikaans was considered a dialect of Dutch. It is now considered a distinct language, for which the WWII and the Boer Wars are at least partly responsible. When the South of Sweden still belonged to Denmark, the local dialect was considered Danish. Now it is considered a dialect of Swedish. Serbo-Croatian was considered a single language with two dialects until the war in Yugoslavia, and ever since Serbian and Croatian are considered to be different.

When a dialect is no longer considered part of a language, this can also drive changes: words dating from before 1920 are usually (about) the same in Afrikaans and Dutch, while 'new' words are more often different: the word for speed bump is verkeersdrempel in Dutch, but spoedwalletjie in Afrikaans (Martin & Gouws, 2000 [787]). Also for Serbian and Croatian, it is (more) likely that new vocabulary will lead to different words in the two languages now that they are considered as different languages. Even more so because Serbian is (mostly) written in Cyrillic, and Croatian in roman script. This role of politics is why it is often claimed that *"a language is a dialect with an army and a navy."*[11].

[10] It should be noted that in Hong Kong (Cantonese dialect) they use simplified characters, making it less intelligible.

[11] This is a saying from the tradition of Yiddish linguistics, and is first quoted by Weinreich: *"A shprakh iz a diyalekt mit an armey un a flot"* (Weinreich, 1945 [13])

Not only can languages and dialects not clearly be told apart, also the difference in their treatment is gradual: national dialects are often treated in many respects as languages. For instance, American and British English are (still) considered dialects of the same language. But in many ways they are more and more treated as separate languages: in 1789 Noah Webster published an article called *"An Essay on a Reformed Mode of Spelling, with dr. Franklin's Arguments on that Subject"* in which he argues that it's more logical to write "analyze" instead of "analyse" because of its pronunciation. This reformed mode of spelling was accepted in America, but not in England, and ever since there is a difference between English and American spelling. And since Webster in 1806 published his first "American Dictionary", there are also special American dictionaries. Similarly, there are special German dictionaries for Austria, and there is constant debate whether there should be a special Dutch dictionary for Belgium. So not only is there a gradual difference between language and dialect, there is also a gradual difference in their treatment: some dialects are treated as languages in (for instance) having their own dictionary.

Since it is possible to treat dialects as languages, the difference between the two is less important for a multilingual system like SIM*u*LLDA: any dialect can be treated as a language when need arises. This might suggest that there is no theoretical problem if all dialects can simply be considered to be languages, but that would not be a valid conclusion. There is no limit to the number of dialects: not only nations and states have their linguistic variation, but also provinces and villages can have their own particular pronunciation and vocabulary. For instance, the words pronounced as /ˈhʏːlbɛzəm/ for vacuum-cleaner, and /ˈlələpiːp/ for telephone tend to be only used in a Dutch village called Groesbeek. And you could even talk about a special 'dialect' for many families, since many families will have their own special words. The logical extreme of this is of course the ideolect, which can be seen as a person's own personal dialect.

So in many ways there is no clear difference between languages and dialects; and ideolect, though solidly definable, is only marginally different from dialect in that it is the 'dialect' of a single individual. So in the landscape of languages, there is a sliding scale from language to ideolect, with dialect in the middle. Even though this is scientifically an unsatisfying situation, it has little actual impact on lexicographic practice, or the SIM*u*LLDA set-up: a language will be a collection of lexemes and any collection of lexemes can in principle be assigned the status of a language. It is up to the lexicographers to decide what they want to demarcate as the language described.

3.2.2 New, Regional, and Infrequent Words

A language in SIMuLLDA is a collection of citation-forms of lexemes. But what citation-forms does the language consist of in this perspective? When does a lexeme belong to a language? For many words, this is a trivial question. But there are marginal cases: new words arise at a daily basis, but when can a new word be counted to belong to a language? How many people have to use the word? This becomes even more difficult with loanwords: when is a foreign word simply a foreign word, and when does it become a loanword incorporated in the language? Also, words fall into disuse, but when does a word no longer count as belonging to a language? Is there a frequency threshold? Given the dialectic variation, which dialect-specific words should the dictionary list?

As these are questions from lexicographic practice (entry selection), it is well described; for instance, Svensén (1993) gives an informative overview. But in lexicographic practice, these questions are answered, not in terms of solid criteria, but in terms of guidelines. The reason is that there are no theoretically sound criteria to define when something should be counted as a word of the language and when not. Each of the questions above involves fuzzy criteria. Let me illustrate this by an example: the matter of new words.

A language is not a rigid object: words come and go. That new words appear is easy to see and understand: new items require new words, so it is only natural that the words computer and helicopter didn't exist a century ago. But words have to come from somewhere, they do not suddenly happen to exist. So someone has to be the first to use it: every language contains words that can be ascribed to a specific author[12]. Normally, new words do not appear for no apparent reason: there has to be a need for the new words, usually because it indicates a new concept, for which there previously was no name, but there are other reasons:

> [D]et är mycket vanligare att ett nytt ord representerar ett nytt begrepp som behöver en beteckning. Det är mindre vanligt att man introducerar en ny term för ett existerande begrepp: medvetna sådana nybildningar sker främst för att undanröja negativa associationer (dårhus blir först sinnessjukhus, därefter mentalsjukhus)[13]. (Svénsen, 1987 [34])

[12] For Dutch, for instance, the cartoonist Marten Toonder (with words like denkraam, and minkukel), and comedians van Kooten & de Bie (jeemig de peemig, kieren) have left their marks on the lexicon.

[13] "It is much the more usual case that a new word represents a new concept in need of a name. It is less common to introduce a new term for an existing concept: such deliberately constructed words arise primarily in order to avoid negative associations (bedlam becomes first lunatic asylum then mental hospital)." [Translation by Svénsen]

Not every use of an invented term leads to a new word, the word also has to 'catch on': if I use the word cleffing for 'being intimate in public' (as a negative word for *lovy dovy*) on a regular basis, people around me will in time understand what I mean by it if I use it often enough. But this does not make it a new English word. The fact that my entire family uses the expression *over zijn frappé zijn* when a wine is over its top (and getting less tasteful with the years rather than more) does not make it a Dutch expression. In many countries, there are many words that are very common amongst students, but are hardly ever heard elsewhere. Do these count as Dutch words? More in general, how many people have to use a word in order to make it a part of the language?[14] Whichever the answer to this question, it will always be an arbitrary frequency threshold, and will never yield a undebatable criterion when to count a word as part of the language. The same kind of fuzziness will result when considering minimum frequency, or necessary recency of a word.

In lexicographic practice, the fundamental question when to count something as a word of the language does not play an important role: the threshold on words is not imposed because less frequent or less wide-spread words are not considered actual words of the language, but because of the limited size of the dictionary. Still, the questions remain more or less the same: how new, frequent, etc. should a word be before it appears in the dictionary? Collins makes heavy use of computer corpora in their dictionary composition, and are in a position to select simply the 40.000 words that appear most frequently in their corpus when compiling a bilingual dictionary of that size; this is also in principle what they do, but afterwards they add less frequent words that are still very useful to have in a bilingual dictionary, such as the names of parts of a car[15]; typically the kind of words you want to find in a dictionary if you're having car problems abroad, but still are not very frequent in normal text. There are also other reasons to deviate from sheer frequency, for instance Verkuyl's criterion of completeness: *"If you include* horse*, then the structure of the chess game domain requires that* king, queen, rook, bishop *and* pawn *also be included."* (Verkuyl, 1994). So if half of the words for chess pieces are above the frequency threshold, and half of them below, either all of them should be incorporated, or all of them should be left out.

In a computational set-up, the situation is different: size is much less relevant, since a single CD can contain up to 1.3 million typed pages of text, which is significantly more than the slightly over 1000 pages a medium-sized dictionary contains. And with technological advances, this number can even rapidly increase: a DVD can already hold up to 10x the infor-

[14]Notice that if the subgroup using the word is homogeneous enough, it will count as a word from a regional/social/professional dialect, more on this in section 5.2.3.

[15]Keynote speech, COMPLEX 2001, Birmingham

mation of a CD. And especially on a DVD, this is simply more than any lexicographer could possibly hope to fill, which effectively means that for digital dictionaries (or lexical databases), there is no limitation on size.

Although this loosening of size limitation has many obvious advantages, it does not solve all problems[16]. When frequency is no longer a criterion, there are no solid criteria left to incorporate words in the dictionary or not. This makes the following remark even more true for SIMuLLDA it being an electronic system: a language-module is a relatively arbitrary collection of lexemes frequently appearing within the corpus of the language, which the lexicographer for contingent reasons decided to incorporate in the system.

To return to the main issue: for many reasons, there are no hard criteria to decide what should count as a word-form (or lexeme) of the language and what should not. So again, the conclusion is that theoretically, a language is little more than an arbitrary collection of lexemes. We can adopt the conclusion by Béjoint:

> What does the dictionary represent? What is the language, where does it begin and where does it end? ... the lexicon of a language does not exist as such, apart from the dictionaries and the minds of the users, and it is impossible to draw a faithful portrait of something that does not exist. (Béjoint, 1994 [183])

3.3 The Interlingual Meanings

Word meanings and concepts play a prominent role in all kinds of disciplines. They are often represented as very richly structured entities, fulfilling all kinds of functions. For instance, concepts are supposed to fix the extension of words. That is to say that according to many, including the denotative version of FCA, concepts should be defined in such a way that it is clear which objects that concept denotes, and conversely that for each object it can be decided whether or not it belongs to the concept. Also, it is often attempted to represent concepts in such a way that they provide all the necessary information for linguistic processes like disambiguation, presupposition, bridging, and anaphoric resolution.

In the SIMuLLDA set-up, interlingual meanings are not such richly structured entities, as will be explained in this section. Subsection 3.3.2 will explain why interlingual meanings are not denotational in nature, subsection 3.3.3 why they are not closely related to cultural differences, and subsection 3.3.4 why pragmatics is not taken as part of interlingual meanings.

[16] It even introduces a problem: with a massive amount of information, the user is prone to receive a lot more information on a search than he finds useful. So a real system has to have some filtering mechanism, providing the user with an appropriate amount of information. Such considerations will be dealt with in chapter 5.4.

The first subsection will focus more directly on the nature on interlingual meanings, around two central questions. The first is: how many meanings does a word have? Or: how fine-grained should we define the notion of meaning? And the second question is: can word meanings be sensibly assumed to be language-independent? To put it differently: can different languages express the same meanings, or has every language its own unique, language-dependent set of meanings?

3.3.1 Homonymy, Polysemy and Metonymy

Distinguishing and counting word meanings is a serious and notoriously difficult problem amongst lexicographers. When considering the definitions of the same word in different dictionaries, the number of senses listed for that word is usually one of the important differences between these definitions. The reason for this is roughly the following: a word can be *homonymous*: one word-form can have various, unrelated meanings. For instance, the word-form sole can mean either the bottom of a shoe or a kind of fish. These are clearly distinct meanings that any dictionary should give.

On the other hand, words can also be used creatively, using the word metaphorically to mean something that is not one of its normal senses. For instance in the case of *metonymy*, the name of a part is used to indicate the whole. An often cited example is the fact that a waitress might use the term ham sandwich to refer to the person who ordered one (*"who's the ham-sandwich?"*)[17]. Also here, there is little doubt amongst lexicographers that this meaning of ham-sandwich should *not* be included in the dictionary.

Apart from homonymy and metonymy there is also *polysemy*: a lexeme is polysemous if it has various, related meanings. Polysemy is in a sense between homonymy and metonymy. Polysemy, like homonymy and metonymy, relates various senses (or meanings) to the same word-form, but these senses are neither as clearly distinct as in the case of homonymy, nor as clearly 'the same' as in the case of metonymy. When considering word-senses, it is even hard to say when they stop being homonymous and start being polysemous, and when they stop being polysemous and start being metonymous. You could say that there is a scale from homonymy to metonymy, with polysemy in the middle. And it is hard to clearly distinguish polysemy from either homonymy or polysemy.

To start with the problem of separating polysemy from homonymy: it is hard to say when two meanings are (or are not) related. There are two criteria for doing so (Lyons, 1995 [58])[18]: two words can be said to be related

[17]This example is after Jackendoff (1990), who attributes it to Nunberg (1979 [149]).

[18]There are other analyses, for instance van Campenhoudt (1994) lists 3 different criteria: syntactic criteria, morphologic criteria and semantico-cognitive criteria.

when they have the same etymological root, or when they are semantically related. Semantic relatedness is a graded notion: every two meanings will have *something* in common (for instance, both meanings of sole indicate a physical object), but some will have more in common than others. In terms of SIM*u*LLDA: most interlingual meanings will share *some* definitional attributes, and semantic relatedness could be defined as a sufficient percentage of shared attributes. However, the exact percentage to count two meanings as related will always be contingent.

And also etymology does not give a hard criterion for a number of reasons. Although two homographs are either derived from the same etymological source or not (thus suggesting a hard criterion), this common past may vary from recent to very distant, resulting in a sort of gradedness. Also, etymological facts are hard to establish and (hence) not always obvious to the layman: are the words for the nails on your hand and the nails you hit with a hammer etymologically related? (yes) And the ears on your head, and the ears of corn? (no). But more importantly, the source can often not be fully established: the reconstruction of the history of a word will always contain a certain amount of guesswork.

In lexicography, the difference between homonymy and polysemy has a direct implication: *"homonyms (or homographs) need separate entries in the dictionary."* (Jackson, 1988 [127]). Also, in relation to homonyms the term *meanings* is used, whereas polysemy is related to *senses*. In SIM*u*LLDA, this situation is slightly different: lexemes do not have separate copies for homonymous terms; lexemes simply relate to a number of interlingual meanings. So lexemes cannot be homonymous, but they can be polysemous. Word-forms, on the other hand, can be homonymous: as we have seen, a word-form like the French juste can appear in (and even be the citation-form of) different lexemes. But word-forms cannot be truly said to be polysemous, since they do not directly relate to interlingual meanings, but only via the lexeme(s) they belong to. So word-forms can be homonymous, and lexemes can be polysemous.

The difference between polysemy and homonymy in SIM*u*LLDA is little more than a terminological issue, and does not have a lot of impact on the system. The difference between polysemy and metonymy, however, will be shown to be more influential. This difference is also hard to establish and a common source for differences between dictionaries, as can be shown by looking at the definitions for rib in LDOCE and COD in table 3.5.

If we label the senses in COD **C1 ... C8**, and the senses in LDOCE **L1 ... L3**, we can observe that there are large differences: **C1** more or less corresponds to **L1**, and **C7** could be seen as corresponding to **L3**. The senses **C5** and **C8** are completely absent from LDOCE. Sense **C3** corresponds to **L2**, but in LDOCE, **C4** and **C6** are mentioned as example of **L2**. And **C2** could be seen as a creative use of **L1**.

The central problem is this: new word senses often start out as creative

rib /rɪb/ *n. & v. -n.* **1** each of the curved bones articulated in pairs to the spine and protecting the thoracic cavity and its organs. **2** a joint of meat from this part of an animal. **3** a ridge or long raised piece often of stronger or thicker material across a surface or through a structure to support or strengthen it. **4** any of a ship's traverse curve timbers forming the framework of the hull. **5** *Knitting* a combination of plain and purl stitches producing a ribbed somewhat elastic fabric. **6** each of the rods supporting the fabric of an umbrella. **7** a vein of a leaf or an insect's wing. **8** *Aeron.* a structural member in an aerofoil.

rib[1] /rɪb/ *n* **1** any of the twelve pairs of bones running round the chest of a person or animal, from the SPINE to where they join at the front **2** a curved piece of wood, metal, etc. used for forming or strengthening a frame: *the ribs of a boat/an umbrella* **3** one of a series of long thin raised lines in a pattern: *the ribs of a leaf*

Table 3.5: COD (top) and LDOCE (bottom) definitions of rib

uses; they are often no more than petrified metaphors[19]. Or they are very common special cases of a more general meaning. Consider rib: the ribs of an umbrella are definitely *'ridges across a surface to strengthen it.'* But this particular kind of strengthening ridge is so common that this sense might be almost as frequent by itself as its general sense. And the fact that for instance Dutch has the general meaning of rib (rib), but has a different word for the ribs of an umbrella (baleinen), strengthens the feeling that it might be a meaning on its own. There is no clear threshold as to how common a sense has to be to count as separate. So the fact that polysemous readings often originate gradually makes polysemy and metonymy (or creative word use in general) hard to tell apart.

Regular Polysemy

There is another, even more troublesome problem with polysemy, that can best be illustrated by the following example: consider the ambiguity between an animal and its meat, such as in the definition for chicken in LDOCE:

chick·en[1] *n* **1** [C] a common farmyard bird. A female chicken is a *hen* and a male chicken is a cock (BrE) / rooster (AmE) **2** [U] the meat of this bird eaten as food.

The problem with this specific case of polysemy is that it is *productive*: not only in the case of chicken can you can talk about the meat by means of the word for the animal, but also in case of turkey, ostrich, even cockroach and basically any other animal you can eat[20]. The main exceptions are those cases where there is a specific word for the meat of the animal, such as veal for lambs, beef for cows, and pork for pigs. The general rule is that you

[19] According to some, *all* meanings are metaphoric in nature.
[20] In spite of the fact that cockroaches are not eaten in most parts of the world.

can use the name of any edible animal for its meat, as long as there is no specific word to indicate the meat itself.

Such predictable sense alternations are often referred to as *regular polysemy*, and the alternation between an animal and its meat is not the only example: Pustejovsky (1995a), describes 7 different types of regular sense extensions, which are listed in table 3.6. To show that these alternations do indeed occur in dictionaries, they are illustrated with examples from the LDOCE dictionary. Notice that the animal/meat polysemy is listed as an example of count/mass polysemy, which also encompasses such examples as *"There was rabbit all over the road."* (Copestake & Briscoe, 1992 [98]).

Count/Mass	**lamb**[1] /læm/ *n* **1** [C] a young sheep **2** [U] the meat of a young sheep - see MEAT (USAGE)
Container/Containee	**bot·tle**[1] /ˈbɒtl‖ˈbɑːtl/ *n* **1** [C] a container of liquids, usu. made of glass or plastic, with a rather narrow neck or mouth, and usu. no handle **2** [C] **(of)** also **bot·tle·ful** /-fʊl/ – the quantity held by a bottle
Figure/Ground	**door** /dɔːʳ/ *n* **1** a movable flat or panelled (PANEL[2]) surface that opens and closes entrances to a building, room, vehicle, or piece of furniture **2** an opening for a door; DOORWAY
Product/Producer	**news·pa·per** /ˈnjuːsˌpeɪpəʳ ‖ˈnuːz-/ *n* **1** [C] also **paper–** set of large folded sheets of paper containing news, articles, advertisements, etc., printed and sold usu. daily or weekly **3** [C] a company which produces a newspaper
Plant/Food	**fig** /fɪg/ *n* **1** [C] (a broad-leaved tree that bears) a soft sweet fruit with many small seeds, growing chiefly in warm countries
Process/Result	**ex·am·i·na·tion** /ɪgˌzæməˈneɪʃən/ *n* **1** (an act of) examining **2** [C] *fml* an exam
Place/People	**cit·y** /ˈsɪti/ *n* **1** a large group of houses and other buildings where people live and work, usu. having a centre of entertainment and business activity. It is usu. larger and more important than a town, and in Britain usu. has a CATHEDRAL **2** [+ *sing./pl. v*] all the people who live in a city

Table 3.6: Pustejovsky's (1995b) Regular Sense Extensions

This so called *lexical conceptual paradigm* of Pustejovsky is not the only theory modelling regular sense extentions: the same kind of meaning alternations have been modelled by means of *lexical functions* (see chapter 5) and by means of *lexical implication rules* (Ostler & Atkins, 1992). These other theories are more elaborate in the sense that there are over 50 lexical functions, and around 130 lexical implication rules. Other examples are the alternation between a transportation device and traveling by means of it (snowboard, ski, step), and an animal and its fur.

These regular sense extensions make the distinction between polysemy and creative word use even more problematic. Firstly, there is necessarily a great amount of arbitrariness which of the animals have their meat

listed as a distinct meaning. Also, regular sense extensions often inherit: the ambiguity between the frame and the glass for window passes over to specific kinds of windows, such as portholes, fanlights, transoms, and rosettes. Being windows, portholes can be painted (in which case the frame is involved), but also broken (in which case it concerns the glass).

The first point is relatively inconsequential: currently, dictionaries do not list a 'meat'-sense for ostrich, but this might change in the future because of the increasing popularity of ostrich meat (in western Europe). The question when it should be listed as a separate meaning, is comparable in a sense to the question whether a low frequency word, like epistemic, should be listed in a dictionary; the only consequence of a positive answer is the enlargement of the dictionary with one additional sense (or word respectively). But the second point has more effect, if we look at it from a different perspective. This can be illustrated by the definitions of porthole and window as presented in table 3.7.

porthole /pɔːthəʊl|| ˈpɔːrt-/ *n* a small usu. circular window in the side of a ship or aircraft

window /wɪndəʊ/ *n* **1** a usu. glass-filled opening in the wall of a building, a vehicle etc., to let in light and air **b** a piece of glass in a window; WINDOWPANE **2** *tech* **a** one of a number of areas into which a computer's SCREEN can be divided, each of which is used to show a particular type of information **b** a part of the Earth's ATMOSPHERE through which radio waves can pass to or from space **c** a short period of time that is the only one that can be used for a particular activity **3** a transparent area on the front of an envelope, through which the address can be seen on the letter inside

Table 3.7: LDOCE definitions for porthole and window

The definition of porthole nicely fits the model of genus et differentiae. However, there is a problem: as observed before (section 2.3), the genus term is intended as a word meaning, and not a word-form. So the definition does not imply that a porthole is *'an area of a computerscreen in the side of a ship'*. But because of the difficulties with polysemy, there is not a clear, single sense of the word window that is meant here: the intended meaning is the combination of **1a** and **1b**; a porthole is a kind of window in *both* of these senses of the word[21], that is to say, porthole is a hyponym of the combination of these two meanings rather than of one of them in particular.

The implication of this is, that the genus term in the definition of porthole refers to a non-existing entity: a meaning that is not as such present in the dictionary. And since the analysis of dictionary definitions in SIM*u*LLDA relies on the lexical definition of the genus term, the analysis will stum-

[21] A definition that nicely illustrates this point is the definition of the Dutch word kratermeer (crater lake, taken from appendix A.3), which is a compound word, and predictably a lake at some specific place (i.e. in a crater). But the GVD defines it as the container: *a crater filled with water*, without any reference to the container/containee ambiguity.

ble over such cases. This failure is not without reason: the absence of the required meaning makes the lexical definition of porthole imprecise, and hence not fit for formal analysis. This problem is not strictly restricted to regular polysemy cases: for various reasons the genus term is often something in between some of the listed word meanings. This is a very important limitation of dictionary definitions, which we will return to at various points throughout this thesis.

In a way, the definition of porthole defines the word-form as having two separate (though strongly related) genus terms: on the one hand *window-1a*, and on the other hand *window-1b*. And since SIM*u*LLDA is a multiple inheritance network (see section 2.1.1), it might seem that this can simply be modelled in the system, in the same way that colt can have two separate hyperonyms foal and stallion[22]. However, this does not work. Subconcepts are defined as sharing all the definitional attributes of their superconcepts. And indeed, colts have both all the characteristics of foals, and all the characteristics of stallions. However, a window is not *'an opening made of glass'* (which is roughly what it would have to be if both a hyponym of opening and piece of glass), but the frame and the glass are two aspects of the complex object in the world denoted by the word window.

The other option is to say that we need to introduce a single, complex meaning for window, which embraces both the frame and the glass. And this is the conclusion that Pustejovsky (1995b) draws from the examples of regular polysemy: we should not stick to a *sense enumerating lexicon*, a lexicon in which all the meanings of a word are nicely enumerated. The two meanings of window should not be treated separately, but as a single, complex definition. One of the reasons he gives is the following: if window were truly polysemous, then in any use of the word in a sentence, either of the meanings has to be selected. But no such selection is necessary: we can say such things as:

(3.2) The freshly painted window was broken.

where 'painted' relates to the frame, and 'broken' to the glass, while both relate to a single occurrence of the word window. So the subject of this sentence is the figure and the ground at the same time, which means that the word window has to have a single meaning that encompasses both the frame and the glass at once.

The solution Pustejovsky proposes (Asher & Pustejovsky, 1998) is a syntactic one: words can be typed as *dot objects*, which basically means that one word can be assigned various ontological types, where the context of the sentence can coerce the word into one of its specific types when desired. It is a way of pushing the ambiguity of window from the lexicon up to the grammar, which allows for sentences like 3.2.

[22] If we take the meaning of stallion where it is not specified as **adult**.

However, since SIM*u*LLDA does not contain a grammar, such a solution is not available. There is a possibly comparable solution: we could allow one word-form to 'simultaneously' relate to two interlingual meanings, or define a 'dot-concatenator' between interlingual meanings, and allow word-forms to relate to such combinations of meanings. However, this does not lead to a satisfactory result unless the system is radically altered. So within the current system, either one of the following solution will have to suffice:

1. porthole has two senses, each with a different genus term, hence mimicking the polysemy of window
2. window has only one of the senses 1a or 1b, where the other is a creative sense extension; the same applies to porthole

Neither of these options is completely satisfactory. In chapter 4, we will discuss whether these options in practice will suffice, when we look at actual lexicographic data.

3.3.2 Word-meaning and Denotation

A great deal of lexical semantic theories deal with word meaning in an extensional fashion; the claim of these theories is that a proper representation of word meaning should reflect the objects belonging to the denotation of that word. There are two sides to this: naming and recognition. Firstly, the representation of the word should capture all the important features of the objects in its denotation[23], and secondly, concepts should be represented in such a way that given an object, it can be recognised as belonging to the concept (or not). In this section, I will try to show that a denotational definition of word meaning is not a viable option, and hence that a dictionary should not aim at such a definition. As a result, the interlingual meanings in SIM*u*LLDA will be defined and interpreted in a non-denotational fashion.

There are two assumptions behind the extensional way of viewing the relation between word and object: firstly, that for every word there is a fixed set of objects in the world that form the extension of that concept. And secondly that these objects share a set of necessary and sufficient conditions that can be used to determine for any given object whether or not it belongs to that concept. A common example is that all bachelors are male and unmarried, and all unmarried male people are bachelors. This also holds for the notion of a concept as defined by the denotative contexts in FCA (section 2.2).

However, both these assumptions have been attacked heavily. The existence of necessary and sufficient conditions (which have been around since

[23]With the Fregean assumption that the denotation of a word is the set of all the objects denotated by that word.

Aristotle, who called them το τι ην ειναι), was attacked (amongst others) by Wittgenstein, who claimed that for games, indicated by the German word Spiel, there are no features that all of them share:

> Betrachte z.B. einmal die Vorgänge, die wir "Spiele" nennen. Ich meine Brettspiele, Kartenspiele, Ballspiele, Kampfspiele, usw. Was ist allen diesen gemeinsam? – Sag nicht: "Es *muß* ihnen etwas gemeinsam sein, sonst hießen sie nicht 'Spiele' " – sondern *schau*, ob ihnen allen etwas gemeinsam ist. – Denn wenn du sie anschaust, wirst du zwar nicht etwas sehen, was *allen* gemeinsam wäre, aber du wirst Ähnlichkeiten, Verwandtschaften, sehen, und zwar eine ganze Reihe. ... Ich kann diese Ähnlichkeiten nicht besser charakterisieren als durch das Wort "Familienähnlichkeiten"; denn, so übergreifen und kreuzen sich die verschiedenen Ähnlichkeiten, die zwischen den Gliedern eine Familie bestehen: Wuchs, Gesichtszüge, Augenfarbe, Gang, Temperament, etc. etc. – Und ich werde sagen: die 'Spiele' bilden eine Familie.[24] (Wittgenstein, 1953 [277])

The claim of many subsequent theories is that this non-existence of necessary and sufficient conditions holds for *all* concepts, even for such classic examples as bachelor.

The existence of a fixed set of objects in the extension of a concept has been attacked by Labov, who did an experiment to show that concepts do not have clear boundaries (Labov, 1973). He showed the objects in figure 3.2 to a group of subjects and asked them to name these objects. According to the subjects, 1 was clearly a cup, 4 a bowl, 9 a vase and 11 a mug. However, there was no clear cut-off point where the objects ceased to be a cup: on the top row from left to right, the things start being less and less like a cup, but there is no fixed endpoint. Hence, there is no clear set in figure 3.2 that is the denotation of cup.

Both these points were taken up by Prototype Theory, as developed by Eleanor Rosch and her followers (Rosch & Mervis, 1973)[25]. In prototype theory, objects belong to a concept whenever they are close enough to the prototypical member of the concept. In principle, this needs only to mean a subtle shift in perspective from the Aristotelian point of view: instead of

[24] Try, for instance, to name the activities that we call "games". I mean board games, card games, ballgames, competitions, etc. What do all of these have in common? – don't say: "They *have* to have something in common, otherwise they wouldn't be called 'games' " – but *look* if they have something in common. – For if you look at them you will not see anything that they *all* have in common, but you will see resemblances, connections, and a whole lot of them. ... I can not describe these resemblances better than by the word "family resemblances"; for the various resemblances cross and intertwine in the same way as members of a family do: build, features, colour of they eyes, gait, temperament, etc. etc. – and I will say: the "games" constitute a family.

[25] An excellent reader on Prototype Theory is "Concepts" (Margolis & Laurence, 1999).

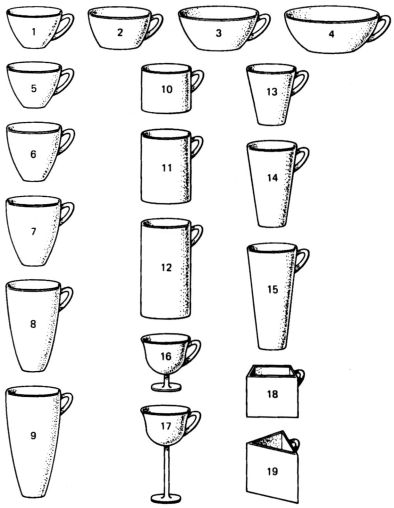


Figure 3.2: Labov's Experiment (Labov, 1973)

sets of necessary features, we have sets of prototypical features, and category membership is a graded notion depending on the number of shared prototypical features. However, the Labov experiment also shows that this could not be a proper solution: between objects 1 and 4 there is no clear set of features that changes, but it is merely a change in shape. So category-membership is dependent on shape and perceptual features, which are notoriously hard to formalise:

> Many critical semantic components may be perceptual and consequently may not be expressible in a spoken language. The general shape of a dog, for example, must be important in defining (in the wide sense) the concept [dog]. However, this shape cannot be described in English to any degree of satisfaction.
> (McNamara & Miller, 1989 [358])

So in many of its incarnations, Prototype Theory has a complex notion of closeness, including perceptual features; the standard example of a prototype used by Rosch is a robin as a prototypical bird. The problem with perceptual features is not only that they cannot be described in English, but also that it is (probably) impossible to formally describe them in any non-perception based framework. So if the extension of concepts (and meanings) critically depends on perceptual features, and perceptual features cannot be embedded in a symbolic framework, then concepts cannot be extensionally defined within a system like SIM*u*LLDA.

That perceptual features are not modelled in SIM*u*LLDA is not to say that they are not important aspects of words; and the same holds for prototypicality, associated ideas, etc. To refer to this rich idea behind a word

I will use the term *concept*; so behind every word-sense is a complex idea that we will not further specify: the concept. The interlingual meanings of SIM*u*LLDA are only a small part of that concept. So SIM*u*LLDA does contain *formal concepts* (which are pairs of interlingual meanings and definitional attributes), but it does not contain concepts. Concepts are related to the denotation of the word-form; interlingual meanings are not. The claim is that this is in correspondence with the lexical definitions in dictionaries, which SIM*u*LLDA aims to model.

There is another problem for the extensional approach to meaning: as shown in the previous section, there is no clear-cut distinction between literary and creative word use. So there is also no clear-cut way to say whether an object that is referred to as a cup truly is a cup, or merely creatively referred to as such. This means that there is no empirical way of establishing the set of cups: it cannot simply be the set of objects called cup often enough, since that set may well contain objects which are only creatively attributed the term cup. This attributes to the non-existence of well-defined extensions for words, and hence the impossibility to define words in terms of that not well-defined extension.

Given these difficulties of denotational semantics, the fact that dictionaries do not try to give a definition that fixes the extension of a word is not a shortcoming of dictionaries, but a necessary evil. For instance, the definition of cup in table 3.8 gives some information to help the reader understand what a cup is, but there are lots of small round containers that are not cups.

cup[1] /kʌp/ *n* **1** [C] a small round container, usu. with a handle, from which liquids are drunk, esp. hot liquids such as tea or coffee

Table 3.8: LDOCE definitions for cup

Not only are interlingual meanings not denotational in nature, they are also not prototypical in nature. As a result of that, no prototypical information should be present as definitional attributes. So, the additions *esp. hot liquids such as tea or coffee* on the LDOCE definition of cup, and *usu. naturally formed* on pool (see appendix A) should be kept outside of the SIM*u*LLDA analysis[26]. However relevant these pieces of information are, they should be modelled differently, together with all the other aspects of the concept. So it will be important to keep concepts and interlingual meanings apart.

[26]There are possible exceptions: if poulain were defined as a *young horse, esp. male*, this would mean that *poulain* has the more general and the more specific meaning (colt and foal) at the same time. Also, the word wadi is defined as a *usu. dry river bed*, which means it is dry most of the year, and not prototypically dry.

3.3.3 Interlingua, Incommensurability, and Cultural Differences

Behind the SIMuLLDA set-up is a strong claim about word-meanings: since words from various languages can express the same interlingual meaning, meanings are interlingual and language-independent. Not all interlingual meanings have to be lexicalised in every language (there can be lexical gaps), but for those interlingual meanings that are lexicalised in more than one language, there is a 'perfect' translation. This language-independence of meaning has been much criticised by semanticists and philosophers. In this section, the question of whether the very existence of interlingual meanings is tenable will be addressed.

Not only does SIMuLLDA allow lexical gaps, it even allows whole ranges of lexicalisation mismatches. According to Sowa (1993), Chinese and English have a very different hierarchy of terms for vehicles, as depicted in the tree in the center of figure 3.3. The way in which SIMuLLDA would cope with such situations is depicted in this figure 3.3[27]: the wordforms in Chinese and English express different (overlapping) sets of structurally related meanings. In this example, there is not a small number of lexical gaps, but many of the interlingual meanings are only lexicalised in one of the two languages. Another well-known example is the word rice, which is hard to translate into Indonesian, since in Indonesia there are many more words for rice (Hutchins & Somers, 1992).

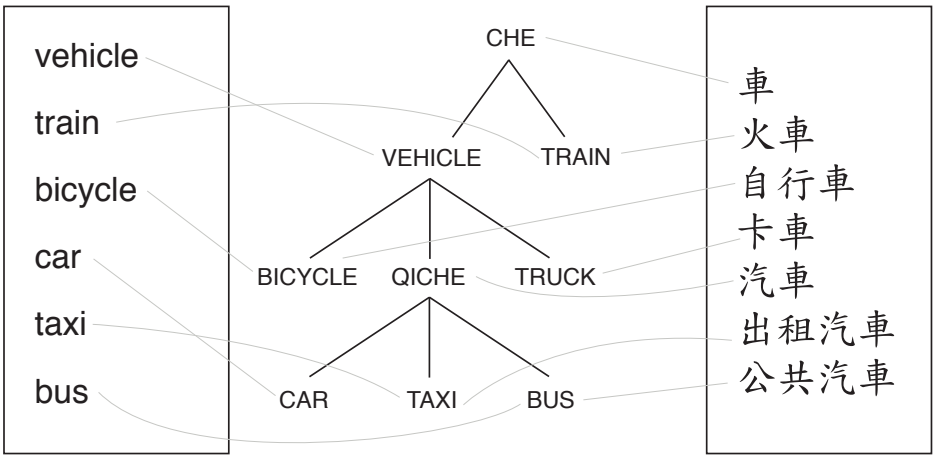


Figure 3.3: Chinese and English Terms for Vehicles (Sowa, 1993)

In principle, SIMuLLDA does not inherently claim a strong overlap in the sets of meanings lexicalised across languages. In fact, SIMuLLDA enforces

[27] Although this hierarchy would not result as such from a FCA analysis of the monolingual Chinese and English dictionaries.

no overlap at all. Theoretically, the relation between two languages and their formal concepts could be as depicted in figure 3.4: if two languages X and Y would use mutually exclusive sets of definitional attributes, the interlingual lattice would consist of two separate parts; one with nodes expressed by language X and the other with nodes lexicalised in language Y, and these two parts would only be connected at the two outer nodes $\top$ and $\bot$. So if two languages are completely incomparable in the sense that all definitional attributes that play a role in the definitions of the words for language X do not play a role for those in language Y and vice versa, there would be no translational synonyms at all between the two languages.

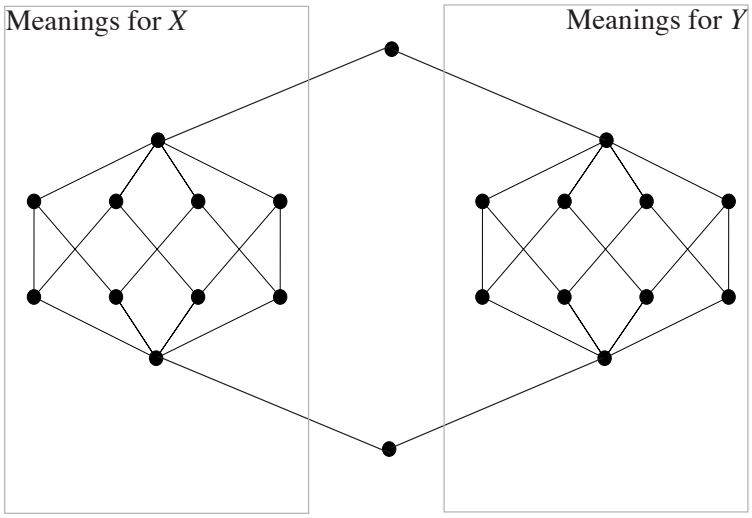


Figure 3.4: Lattice for 2 Unrelated Languages

If languages would behave like this, the SIM*u*LLDA approach would not be very fruitful: every word would form a lexical gap, and the lexical gap filling procedure would not fail theoretically, but clearly practically since the only translatable hyperonym for any word would be $\top$. Also, in this fashion the meanings cannot be truly said to be interlingual, since every meaning except for $\top$ and $\bot$, would relate to word-forms of only one language. So the SIM*u*LLDA setup only makes sense if there is sufficient overlap between the various languages.

On the other hand, the SIM*u*LLDA setup would also not be very sensible if no lexical gaps would exist, for in that case a much easier system of cross-language word-sense identification would suffice (see section 1.2.2). So the system presupposes languages to have a large but incomplete overlap in lexicalised meanings, with a relatively small number of lexical gaps. That there are lexical gaps will not be much disputed, so the question is: is it a

reasonable assumption to claim that languages have a large overlap in their lexicalised meanings?

Incommensurability & Translatability

Within the philosophy of language some leading philosophers, among whom Kuhn (1970), Feyerabend (1962), Quine (1960), and Whorf (1956), have argued that it is fundamentally impossible to translate from one language in another. Feyerabend calls this the *incommensurability* of languages, and it would imply that all meanings would be non-corresponding across languages. The idea is roughly the following: language plays a very prominent role in our cognition, and it is more than just a technique of expression: *"language produces an organization of experience."* (Whorf, 1936 [55]). Language is not just a way of expressing things about the world, it attributes structure and meaning to the world: *"the mind and the world jointly make up the mind and the world"* (Putnam, 1981 [134]). So sentences are always embedded in a particular view on the world, or a *conceptual scheme*. Moreover: sentences only make sense *within* that conceptual scheme. Because of this dependency, a sentence in one conceptual scheme cannot be transferred into another.

There is a strict relation between language, conceptual scheme and translatability: *"where conceptual schemes differ, so do languages. But speakers of different languages may share a conceptual scheme provided there is a way of translating one language into the other."* (Davidson, 1974 [184]). So the prominent position is not that different languages are by definition untranslatable, but that they can be untranslatable, and are so when they are based upon a different conceptual scheme. A well-known example is Kuhn's exposition that Hopi is too alien w.r.t. English to be translatable.

Incommensurability is far from an undisputed position; there are as many philosophers arguing against it as there are defending it. Probably the most famous attack is found in Davidson's *"On the Very Idea of a Conceptual Scheme"* (Davidson, 1974), who argues that the very notion of incommensurability is incoherent. The part of his argumentation that is most relevant for the current discussion is the following: the idea behind a conceptual scheme is that you can have different points of view for looking at the world. But that presupposes that there is a constant thing that you can have different viewpoints *on*, for otherwise the viewpoints would be too incomparable to call them different. The only way to make sense of different viewpoints is by means of language: two people have different conceptual schemes if they speak languages that are not intertranslatable. This failure can never be complete: translatability is a criterion of languagehood, so if something cannot be translated at all, it is not a language. And partial failure of intertranslatability critically depends on the mistaken idea that there are two separable things: a world to be organised and the organisation of

the world. This mistaken duality between scheme and content is what he calls the *"Third Dogma of Empiricism"* (Davidson, 1974 [189]).

This has led Quine to admit that (in)commensurability and translatability are not the same notion[28]: *"Translatability is a flimsy notion, unfit to bear the weight of the theories of cultural incommensurability that Davidson effectively and justly criticizes."* (Quine, 1981 [42]). If this is the case, than no matter how interesting the notion of incommensurability is, it is not directly relevant for the current thesis, since the prime concern here is the lexical translatability.

Cultural Differences & Intertranslatability

Davidson has shown that untranslatability is not a good criterion for incommensurability. But that does not mean that cultural incommensurability does not entail untranslatability. To put it in more concrete terms: (culture-dependent) words are imbedded in a cultural setting; does this not simply entail that all words that are culture-dependent are untranslatable? We will look at this problem from a less philosophical and more practical point of view, by looking at some culture-dependent concepts and the translatability of the their lexicalisations.

A good example of a culture-dependent phenomenon is the difference between a lunch in Italy and a lunch in the Netherlands. In Italy, lunch (colazione) is a warm meal, most commonly pasta. In the Netherlands, lunch (middagmaal/lunch) is a couple of sandwiches and a cup of coffee. The difference between the prototypical lunch in the two countries is so big that it is not uncommon to hear Dutch tourists in Italy (and vice versa) utter: "This is no lunch!". Since concepts depend on prototypes, this might be taken as implying that the Italian word colazione and the Dutch word middagmaal do not share a common meaning.

However valid this may seem, consider a lunch in England and in Singapore. The difference between these two is easily as big as the difference between a Dutch and an Italian lunch. So we should conclude that the respective terms for 'lunch' in the two languages should not be translatable. But the problem is that in both countries they speak the same language: English.

The conclusion from this should be that culture-dependence is not a good criterion for translatability, since there is no strong overlap between language and culture: the same language is often (for historic reasons) spoken by people of entirely different cultures[29], and in multilingual countries (often) different languages are spoken by people of the same culture.

[28]Though Kuhn and Feyerabend simply introduced the term *incommensurable* as meaning *not intertranslatable*.

[29]Quine's incommensurability argument also operates within one language.

A possible answer is that not only languages, but also dialects can result in conceptual schemata, the concepts of which would be uninterpretable outside of it. Although this would solve part of the problem, it is far from an elegant solution. And given many gradual changes in culture, and the huge number of dialects, it does not lead to a workable situation. Furthermore, as we have seen in section 3.2, there is a sliding scale from language, via dialect to ideolect. And there also is not a limited number of dialects. So since there is no clear extension of the notions of a language and a dialect, one can hardly claim that words are only interpretable within the boundaries of a language, or within the boundaries of a dialect.

The lunch-example sketched above is based on a denotational argument: the words in the two languages are supposed not to have the same meaning because their extension is different. But as seen in section 3.3.2, SIM*u*LLDA uses a non-denotation-based notion of word-meaning. So in a way, the example does not show much; the best we can say is that the differences in lunch-culture do not conclusively show that the words colazione and middagmaal do *not* share a common meaning. Nonetheless, there are clear differences between the two kinds of lunch. So in what way can we truthfully claim that they *do* share a common meaning?

Given the fact that SIM*u*LLDA is based upon the analysis of the content of monolingual dictionaries, the answer should lie in the dictionary definition for the different languages (from Dvd, Garzanti and LDOCE resp.), as shown in table 3.9.

ontbijt', o. (-en), **1**. eerste maal van de dag.	**ontbijt'** first meal of te day
mid'dagmaal, o. (...malen), maal dat men 's middags gebruikt.	**mid'dagmaal** meal eaten during the afternoon
colazione [-zió-] *s.f.* pasto del primo mattino o di mezzogiorno.	**colazione** meal of the early morning or afternoon
cena [cé-] *s.f.* il pasto della sera.	**cena** meal of the evening
break·fast /ˈbrekfəst/ *n* the first meal of the day.	
lunch /lʌntʃ/ *n* a usu. light meal eaten in the middle of the day.	
din·ner /ˈdɪnəˣ/ *n* the main meal of the day, eaten either at midday or in the evening.	

Table 3.9: Words in Different Languages for Daily Courses

All the different dictionaries give a very modest definition of the various courses; the LDOCE entry for breakfast, for instance, makes no mention of bacon and eggs, or anything else particular for the English breakfast culture. It simply states that breakfast is the first meal of the day. This, I claim, is not a shortcoming: the English word breakfast does not mean the English breakfast in particular, but simply *'the first meal of the day'*, just like

the Dutch word ontbijt does. So despite the cultural differences between England and Holland in this respect, the words ontbijt and breakfast have the same meaning. Of course the concept behind it, and the related prototype, will be different. But this is not relevant for the intended notion of an interlingual meaning. Notice also that the words colazione and middagmaal *are* different, but not because of their different extensions: collazione is a more general term, encompassing both breakfast (prima collazione), and lunch (pranzo).

The example above shows that cultural differences do not, by definition, result in a difference in word-meaning. But the fact that it does not necessarily lead to differences in word-meaning does not entail that it cannot lead to such differences. In fact, the example could very well be a relatively isolated case in which there is cross-language overlap in meaning despite the cultural differences. In order to assure that there is a large overlap in meaning between languages, we need a more compelling argument; but to my knowledge there are no definite arguments why languages should have a large overlap in meaning. The only thing we can observe is that we all share the same habitat, and that the stability of the world *might* ensure a relative uniformity in the classification of the world. So the best argument in favour of a large overlap is the kind of argument Martin gives:

> Certaines données du monde, physiques, physiologiques, anthropoculturelles, exercent sur la vie des hommes une si forte contrainte qu'il est impossible qu'elles ne laissent aucune trace dans la langue. Et ces traces, du fait même, ont toute chance d'être des universaux.[30]
> (Martin, 1983 [88])

Such arguments, however, can never be conclusive. In fact, if we do not take coextensiveness as a hard criterion for correctness of translation (as I have argued we should not), it is has to see what could guarantee that a given translation is correct. The simple fact that two words have similar or identical dictionary entries can never be a philosophically satisfying argument for their sameness of meaning. However, as said before, the goal of this thesis is not to improve upon the content of dictionaries, but to accept them as correct semantic characterisations. And lexicographers are well aware of the problems with translatability, but opt for the practical solution: *"Practiciens, les lexicographes partent évidemment de la possibilité de passer d'une langue à une autre langue, tout en reconnaissant ses limites."*[31] (Messelaar, 1990 [18]).

[30]Certain physical, physiological, anthropocultural facts about the world impose such a strong constraint on the life of human beings that it is impossible that they should not leave a trace in the language. And these traces, by themselves, have all the chance of being universal. [My translation]

[31]Practicians, lexicographers evidently part from the possibility to pass from one language to another, fully recognising its limitations.

If we take dictionary definitions seriously, the amount of overlap between languages is an empirical question: the fact that the MultiWordNet project found 5% lexical gaps in the Collins English-Italian dictionary means that about 95% of all meanings should coincide between English and Italian. So the overlap in meanings between the various languages of the world is simply not an a priori fact, but an empirical claim that can be falsified. It is this empirical claim, amongst others, that will be put to the test in chapter 4.

3.3.4 Word meaning and the Colour of the Word

Word-forms are synonymous in SIM*u*LLDA if they express the same interlingual meaning. But words that are synonymous can still differ significantly. This is illustrated by the the list of words in table 3.10, all of which indicate a policeman.

bobby *infml BrE* a policeman
bull[1] **3** *sl, esp. AmE* a policeman
cop[2] *infml* a policeman
copper[4] *infml* a policeman
flatfoot *sl* a policeman
peeler[2] *BrE old sl* a policeman
pig[1] **4** *derog sl* a policeman

Table 3.10: Lexical entries for 'policemen' (Vossen, 1993 [252])

The difference between these synonyms is that they all have a different applicability: in contexts where the one is perfectly feasible, the other is not. Hence any good dictionary, and especially a bilingual dictionary, should indicate and explain the difference between them. On the one hand, the non-native person producing an English text should be warned that pig is not to be used in official texts, and on the other hand, it should be indicated that the English word pig would be more appropriately translated by smeris, kit, or klabak (all informal words for policemen) in Dutch than by politieagent (the neutral term).

This means that a multilingual system like SIM*u*LLDA should incorporate labels in one form or another. They can be very straightfowardly incorporated: the label *infml* could be taken as just another definitional attribute of copper, which would map it onto the Dutch word smeris, having a similar definition (*agent van politie*), and a similar label (*inform.* in DvD).

This would, however, not lead to a methodologically correct solution: given the characterisation of the various parts of the system, such a solution would imply that a flatfoot is a special *kind of* policeman, which it is not:

the word flatfoot is an informal *word* to indicate a policeman, not a word for an informal policeman[32].

However, there are two problems with this: firstly, the fact that pig and policeman refer to the same objects can not be (directly) used as an argument to distinguish meaning-related features from word-related features since, as discussed in 3.3.2, SIM*u*LLDA does not take a denotational stance towards word meaning. Secondly, it is also not the case that labels belong to the word-forms themselves: it is not the word pig that is informal, since it is a perfectly normal way to indicate the animals: the informality is related to the word-sense.

In the SIM*u*LLDA set-up, labels will be therefore related neither to the word-forms nor to the interlingual meanings, but to the 'thing in the middle': the word-sense. As already said, the actual implementation will be discussed in section 5.2.3. For now, the only relevant thing is to observe that interlingual meanings are not marked with usage information: if two words link to the same interlingual meaning, this means that they share the same content, not that they are strictly synonymous in the sense that they can also be used in similar contexts, or are directly useable as an appropriate translation.

3.4 The Definitional Attributes

The definitional attributes in SIM*u*LLDA are the semantic features that define the interlingual meanings. Since interlingual meanings were characterised in the previous section as little more than sets of definitional attributes, these play are crucial role in the SIM*u*LLDA set-up. And this might seem a disturbing property of the system, since in that light, definitional attributes resemble the much despised semantic primitives from generative semantics. In this section, I will argue that definitional attributes are much more like the *sèmes* of the French structuralism, first described by Hjelmslev (1959). After that, some properties of definitional attributes will be reviewed.

3.4.1 Sèmes, Semantic Primitives, and Interpretative Semantics

In order to give (lexical) semantics a solid foundation, a number of theories have been proposed in which there are atomic semantic elements, or semantic primitives. The most widely known of these is the *structural semantics* theory by Katz and Fodor (1963). They propose a system with building

[32] As will be discussed in section 5.2.3, this is not true for all labels: perdu is an old word for a soldier assigned to dangerous duty, but mace is a word for an old weapon. These should clearly be distinguished.

block of meanings that they call semantic markers: *"Semantic markers represent the conceptual constituents of senses in the same way in which phrase markers represent the syntactic constituents of sentences."* (Katz, 1972 [140]).

These semantic markers provide the foundation of meaning, and are equipped with a lot of strong properties: knowledge about semantic markers is innate, so semantic markers do not have to be learned. Also, since they are related directly to the brain, they are independent of a specific language, which means that every language should have the same semantic markers. Ergo, semantic markers are universal. Also, they are indivisible and undefinable: they are innate elements that are supposed not to *need* definition. Like physical atoms, phonemes and syntactic markers, there is supposed to exist only a limited number of them.

Structural semantics has been heavily criticised, and effectively rejected, most prominently by Vermazen (1967), Partee (1975), and Lewis (1972). The fiercest criticism is the *grounding problem*: structural semantics fails to articulate the relation between semantic markers and the world; it does not explain how these semantic building blocks are grounded in reality. Because of this problem, structural semantics fails to provide an account of the conditions under which expressions are true (or false). And expecially in the seventies, truth conditions were of crucial importance: *"Semantics without truth conditions is no semantics"* (Lewis, 1972 [170]). Without grounding, the semantic markers themselves are again in need of interpretation; giving a componential analysis in terms of semantic markers does not explain anything, but is merely a translation from English into 'Markerese'.

A possible way out would be to say that semantic markers do in fact determine their denotation. That would solve the grounding problem, but lead to a different kind of difficulty: if semantic markers are related to the external world, it is hard to still maintain that they are innate: the external world is definitely not innate, so somehow it should be explained how semantic markers get this grounding. And if they receive their grounding in reality by means of sensory input and interaction with reality, then at least the groundedness of semantic markers is no longer innate.

But even without denotational problems, the idea of innateness of semantic markers is problematic: innateness should entail that semantic decomposition and definitions are psychologically real. But psychological experiments have consistently shown that people do not behave as if they have definitions. For instance, if the concept *filly* would be a mental complex of *foal* and *female*, then *filly* should be 'harder', i.e. take more processing time and generate slower response times, than *foal*. However, such results have never shown up in practice[33].

As an additional point, it is hard to actually find semantic analyses in

[33]In this particular example, *filly* will even produce the slower response times, given salience and frequency effects.

terms of semantic markers. Even Fodor himself (1998) rejects its practical feasibility:

> There are practically no defensible examples of definitions; for all the examples we've got, practically all words (/concepts) are undefinable. And, of course, if a word (/concept) doesn't have a definition, then its definition can't be its meaning. (Fodor, 1998 [45])

Another problem is the number and the exact set of semantic markers: there seems to be no way to tell which the semantic markers should be. It is often taken by critics as a case in point that Wierzbicka, who has made the best attempt at actually describing semantics using semantic primitives (Wierzbicka, 1972; Wierzbicka, 1980), has often changed her proposed set of semantic primitives[34].

Nowadays, the terms 'Markerese' and 'generative semantics' are used more like a term as 'boogie-man', to indicate that a semantic theory has gone astray.

Interpretative Semantics

All this heavy criticism has not silenced the idea of semantic primitives. Within the tradition of the French structuralism, the idea of componential semantics is still very much alive. The theory in which they appear is called *interpretative semantics* (sémantique interpretative), and the difference with Katz & Fodor style semantics is that the French version of primitives is much more modest. It is this more modest version of semantic primitives that is most relevant for this thesis.

In interpretative semantics, the semantic primitives are called *sèmes*, and the meanings they constitute are called *sémèmes*. Sèmes come in two versions: *sèmes génériques*, that label a sémème as belonging to a certain class, and *sèmes spécifiques*, that distinguish the sémèmes within the same class. Prominent scientists in the tradition of interpretative semantics are Greimas, Pottier, and Rastier. Rastier explicitly rejects the central claims of structural semantics (Rastier, 1987): according to him, sèmes are not universal, not (interestingly) indivisible, they are not (necessarily) small in number, and they are not qualities of a referent or part of a concept. This much more modest approach makes sèmes much more like definitional attributes than semantic markers are, since definitional attributes are also not denotationally determined (as discussed in section 3.3.2), and not (entirely) universal since the existence of lexical gaps implies that certain definitional attributes might not used for any lexicalisation in a language (as discussed in section 3.3.3).

[34] One could argue that Wierzbicka has empirically determined a workable set, although this decreases the feel of principled primitiveness.

The basic conception behind sèmes is in many ways close to that of definitional attributes: sèmes are features that define a meaning, and they are often even linked to dictionary definitions. The resemblance is most striking in an example by Messelaar (1990), who describes the French words for cow and lamb in term of the sèmes **adult**, **non-adult**, **male** and **female** (see figure 3.5, which is remarkable similarity to table 2.5).

sèmes / lexèmes	bovin	adulte	non-adulte	mâle	femelle
vache	+	+			+
veau	+		+		

Figure 3.5: Analysis of Sémèmes with Sèmes (Messelaar, 1990 [69])

Messelaar even proposes to use the system of sèmes to compare words across languages, and compare translational synonyms by comparing their sèmes. This comparison then gives the number of sèmes on which the two heteronyms differ (see quote below), which is rather close to the basic principle behind lexical gap filling in SIM*u*LLDA (section 2.3.2).

> [P]our nous il s'agit surtout de chercher des couples de mots dont la différence sémantique semble minimale, et de comparer leurs sémèmes (présence et absence des sèmes) en vue de l'admission ou du rejet des équivalents possibles ... l'analyse sémique est une technique permettant de montrer le degré d'équivalence sémantique entre deux hétéronymes.[35] (Messelaar, 1990 [69])

However striking these resemblances are, this does not mean that sèmes are identical to definitional attributes. There are important (methodological) differences, which will be discussed shortly. But given the resemblances, much of the discussion around sèmes is highly relevant for a better understanding of the nature of definitional attributes. Let me first reproduce Rastier's arguments against the central claims of semantic markers.

[35]For us, it involves mainly searching pairs of words for which the semantic differences seem minimal, and comparing their sémèmes (presence and absence of sèmes) in the light of the admission or rejection of possible synonyms ... semic analysis is a technique that allows to show the level of equivalence between two heteronyms.

Resemblances

As already mentioned, Rastier rejects all the common properties of semantic markers mentioned at the beginning of this section. According to Rastier, sèmes should not be considered picking out their extension, since a definitional approach to meaning leads to 'insurmountable problems', some of which were discussed in section 3.3.2. But Rastier adds a fundamental issue to these problems: according to denotational semantics, the naming and categorisation of objects is the main objective of semantics. Now according to Rastier, this is a fundamental mistake: language is not a simple nomenclature for assigning names to objects, but a mechanism for communication. Denotational semantics can only yield a 'linguistique de signe', and never a complete interpretation of text. Interestingly, in the continental tradition prototype theory is viewed as a kind of denotational semantics, and hence falls prey to the same arguments:

> La notion de sens prototypique est l'aboutissement de cette tradition qui donne le primat à l'ordre référentiel, quand elle suppose un objet prototypique représenté par un concept prototypique. Par ce primat, la méthode sémasiologique définit les mots par les choses, et maintient l'illusion archaïque que la langue est une nomenclature.[36] (Rastier *et al.*, 1994)

According to Rastier, sèmes do not have a special status: there is nothing that makes them universal, nor are they miraculously economic and hence small in number. They are just empirical tools: *"Le sème est le trait distinctif sémantique d'un sémème, relativement à un petit ensemble de termes réellement disponible et vraisemblablement utilibables chez le locuteur dans une circonstance donnée de communication."*[37](Pottier, 1980). So sèmes serve a practical purpose (of distinguishing meanings), and their number and identity has to be experimentally established. This means that differences in sèmes between languages are to be expected:

> [L]'opposition sémique /intra-urbain/ vs /extra-urbain/ n'existe en français que parce-qu'elle permet de distinguer des sémèmes comme 'train' et 'metro', 'route' et 'rue', 'autocar' et 'autobus'. Et il est fort douteux que cette opposition sémique existe dans les langues ama-

[36]The notion of a prototypical meaning is the result of that tradition which gives precedence to the referential ordering, because it assumes that a prototypical object be represented by a prototypical concept. By this precedence the semasiological method defines words by means of objects, and maintains the archaic illusion that language is a nomenclature.

[37]The sème is the distinctive semantic trait of a sémème, relative to a small group of terms truly available and convincingly useable by the speaker in a circumstance arising in communication.

> zoniennes, par exemple.[38] (Rastier, 1987 [28])

So in the tradition of interpretative semantics, semantic primitives (sèmes) are much less burdened with properties than in the Katz & Fodor proposal. They are not supposed to define meaning independently, but to help the human speaker/hearer in the interpretation of text. In this, they are supposed to have properties that do not apply to definitional attributes, or even semantic markers.

Differences

Despite their similarities, sèmes and definitional attributes are not the same on many accounts. The reason is that they are supposed to serve an entirely different function: definitional attributes help to order words in lexicographic databases, while sèmes represent semantic information helpful in the interpretation of texts.

The main function of sèmes spécifique is to distinguish and oppose the sémèmes within the same taxème. But opposing sémèmes just by their having or not-having a certain sème is, according to interpretative semantics, not desirable. The reason for this is that such a method of opposing would not be bound to the sémèmes of the same genus term; it would oppose the sémème to *all* other sèmemes that do not have the sème in question:

> L'universalité de l'opposition ainsi traduite: en quelque sorte, 'couteau' se trouvera opposé à tout sémème, du même taxème ou non, qui ne comporte pas le sème /pour couper/. [39] (Tanguy, 1997 [58])

In SIMuLLDA, definitional attributes do thus oppose a formal concept to all other an do so by design: opposing a formal concept to all others is not just seen as non-problematic, but actually as an attractive feature. Without this feature, there would be no multiple inheritance, and no lexical gap filling procedure.

The sémème is supposed not just to capture the formal meaning of the word, but at least part of the idea behind it. This becomes clear by the following quote from Rastier: *"Le Petit Larousse définit ainsi* caviar*: Œufs d'esturgeon salé. Ce type de définition nous paraît insuffisant, car le trait /luxueuse/ devrait y figurer."*[40] (Rastier, 1987 [63]). The reason he gives is that

[38] The semic opposition /intra-urban/ vs /extra-urban/ only exists in French because it allows the distinction of sémèmes like 'train' and 'underground', 'street' and 'road', 'bus' and 'coach'. And it is very unlikely that this semic opposition exists, for instance, in the amazonian languages.

[39] The universality of the opposition thus translated: in some sense, knife will be opposed to all sémèmes, whether or not belonging to the same taxème, that do not have the sème /for cutting/.

[40] The Petit Larousse defines caviar as: "Salted eggs of a sturgeon". This type of definition seems insufficient, since the element /luxurious/ should appear in it.

22 out of 28 students he questioned named 'luxuous' in their definition of caviar.

The difference between sémèmes and definitional attributes is nicely illustrated by the fact that there are also virtual sémèmes, which are related associations to the sème, that help to disambiguate words, and interpret the word in context; virtual and normal, sémèmes can even be interchanged in special circumstances.

There is also one aspect in which definitional attributes are stronger than sèmes: sèmes do not have an independent status: they only serve to distinguish sémèmes within the same field (having the same genus). Definitional attributes, however, are supposed to be independent of the genus term. Besides the fundamental issues this raises, it also leads to a direct practical problem: some attributes are directly dependent on their genus term. These will be discussed in the next section.

3.4.2 Lexicalisation of Definitional Attributes

Like interlingual meanings, definitional attributes can be lexicalised in a language. And given the language-independence of the interlingual concept lattice (see section 2.3.1), this lexicalisation will be located outside of the interlingua, in the language boxes.

In this, the relation between languages and definitional attributes is comparable to the relation between languages and interlingual meanings. But there are two important differences. The first difference concerns lexical gaps. By the definition of lexical gaps on page 43, it is possible for an interlingual meaning not to be lexicalised in some language.

Lexical gaps can be filled by the process of lexical gap filling, by taking the nearest lexicalised superconcept, together with the definitional surplus. Many definitional attributes are subordinate attributes (see page 51): attributes that are only appear in the context of some concept. This because many attributes are used only as a differentia specifica for a specific genus term and nowhere else. Now if the definitional surplus contains such a subordinate attribute, then that attribute will not play a role in the definitions of the language containing the lexical gap.

For the lexical gap filling process to function properly, it is necessary that those definitional attributes that do only play a role in other languages are nonetheless lexicalised. So for definitional attributes we will require that every single one of them is lexicalised in every language within the system.

The second difference is the following: interlingual meanings are lexicalised by means of lexemes, and linked to the interlingual meaning by means of their citation-form.

3.4.3 Adjusted Attributes

The discussion of attributes thus far did not take into account some of the slightly problematic definitions that are found in dictionary practice. In this section, I will discuss some types of problematic definitions, explain why they pose a problem, and make slight adjustments to the system where necessary and possible.

Dependent Attributes

As discussed in section 3.4.1, definitional attributes are, contrary to sèmes spécifiques, global in the following sense: an interlingual meaning that has a certain attribute is opposed to any meaning that does not have it, and not just to other hyponyms of the same superconcept that do not have it. This directly implies that the definitional attribute is to carry its meaning without the presence of the genus term.

At least in some cases, this leads to problems, for instance in the definition in table 3.8, where cup is defined as *"a small container"*. The definitional attribute **small** cannot be viewed independently of the genus term: a cup is not 'small' in any arbitrary sense of the word, it is *small for a container* (as is well-known by now)[41]. In a sense, in every definition that uses the characterisation *small* as part of its definition, a different property is indicated. There are four ways to deal with this:

1. one could say that such dependent descriptions are a flaw of paper dictionaries, and disallow them in SIM*u*LLDA
2. one could simply accept that it is a different attribute on every occasion and extend all attributes of this kind with their restrictor, i.e. interpret *small* in *a small container* as expressing the definitional attribute *small for a container*.
3. one could simply ignore the problem and have all words with **small** in their definition slightly incorrectly share a common superconcept and have the user interpret it in the correct way
4. one could extend the system by allowing restrictors as functional parts of definitional attributes.

The last of these options is by far the most elegant (although the others might suffice). The way it works is very simple, and just a slight variation on the second option: definitional attributes can take the form **small**

[41]Not everything that traditionally is seen as a dependent attribute is so in SIM*u*LLDA: given that rabbits have a life span of 8-10 years, an 11 year old rabbit is very old, while an 11 year old boy is still young. Nonetheless, **young** is not dependent, since it is interpreted as 'not yet sexually mature'.

rest(container). This will be a different attribute than any other version of **small**, but with a special property: normally, the attribute will be lexicalised as *small for a container*. Whenever it is used in combination with its own restrictor, it can simply be rendered as *small*. So for the interlingua, this is identical to the second option: in combination with every other genus term, *small* expresses a different definitional attribute. But the lexicalisation of these definitional attributes is adepted. The actual effect of this will be discussed in the next chapter.

Culture Specific Attributes

Some concepts are very culture specific, and hence their definitions will also be. Consider the Dutch word Sinterklaas[42]:

Sin·ter·'klaas (de ∼ (m.)) **1** heilige die tegen of op zijn naamdag aan kinderen geschenken brengt ⇒ *Sint-Nicolaas*, *Sint* **2** naamdag van St.-Nicolaas (6 december) of het huiselijk feest op de vooravond daarvan, waarop anoniem geschenken worden gewisseld	**Sin·ter·'klaas** **1** saint that gives presents to children at or around his name day **2** name day of St. Nicholas (december 6) or the homely festivity on the preceeding evening, on which presents are exchanged anonymously

In its second meaning, Sinterklaas is said to be a name day celebrating *Sint Nicolaas*. This can be analysed in term of the definitional attibutes of name day (which we will ignore here), and the definitional attribute **van St.-Nicolaas**. The issue is that this definitional attribute is itself culture specific: although Santa Claus is named after Sint Nicolaas (a legend based upon the Bisshop of Myra), the character of Santa Claus is more closely related to the Russian legend of *Дед Мороз* (Grandfather Frost); Sinterklaas is a very specifically Dutch tradition. So the definitional attribute **van St.-Nicolaas** is maximally restricted in its application: it will only appear in the scope of name day, and it will only relate to precisely one word-form in the Dutch language and to no other languages.

This example raises three questions. Firstly: does it not indicate that there is an incredible number of definitional attributes, resulting in an uncontrollable number of formal concepts? Secondly: does it not mean that language are in fact incommensurable, as discussed in the previous chapter? And thirdly: since definitional attributes have to be lexicalised in every language in order for lexical gap filling to work, how can the definitional attribute **van St.-Nicolaas** be fruitfully lexicalised in a language that does not have the notion of Sinterklaas?

[42]This example was provided by Eddy Ruys

The first two are easy to answer: hopefully, the number of culture specific attributes is limited (which is an empirical question), so the number of definitional attributes will stay manageable; but even if there are many, that will not largely affect the number of formal concepts, since as discussed in section 2.4.6, restricted attributes only yield a very limited number of additional concepts. Also, a single example does not result in incommensurability; this specific example only shows that there are also words for which there is a lexical gap in *every* other language.

The third problem is more serious and no truly satisfactory solution exists. This is not too much a defect of SIMuLLDA: bilingual dictionaries face the same problem. If we look for instance at the van Dale N-E definition, no actual translation of the second meaning of Sinterklaas is given:

> **sinterklaas 0.1** [(persoon verkleed als)] Sint-Nicolaas] *(St. Nicholas)* ⇒ ± *Santa (Claus);* ⟨vnl. BE⟩± *Father Christmas* **0.2** [feest] ⟨*feast of St. Nicholas*⟩

So the solution is to simply lexicalise the definitional attribute **van St.-Nicolaas** as `of St_ Nicholas` in English, even though this is not particularly helpful[43]. If we lexicalise **van St.-Nicolaas** in this fashion, SIMuLLDA can generate the kind of definition in the VDNE dictionary: The lexicalisation in English of the superconcept **NAME_DAY** will be name day, and the lexicalisation of the definitional surplus will be `of St_ Nicholas`. So the complete definition will be: *"name day of St. Nicholas"*.

3.4.4 The Value of Dictionary Definitions

As has been mentioned at various places, SIMuLLDA is able to model the content of monolingual dictionaries in such a way that it can yield bilingual dictionaries. At the core of the system are the definitional attributes, and definitional attributes in turn are little more than unravelled dictionary definitions.

So this thesis is founded on the assumption that definitions given in dictionaries are in principle sound and useful. However, dictionary definitions have been much criticised over the last few decades. There have been all sorts of fundamental objections, a good summary of which is given by Béjoint:

> Most linguists writing about general-purpose dictionaries since the fifties have been dissatisfied with the definitions: they do not take into account folk definitions (Weinreich, 1962a [30]), they are circular (Wierzbicka, 1980 [81]), they mix up information about the world and information about the sign (Weinreich, 1964), they do not indicate

[43] Notice that an additional unsatisfactory element of this solution is, that this definitional attribute will not be linked to the word *St. Nicholas* or Sinterklaas present elsewhere in the system.

> obligatory syntactic patterns (McCawley, 1973 [167]), they use terms that are not kept constant throughout the dictionary (Gleason, 1962 [100]), they do not distinguish between meaning and context-bound 'application' (Weinreich, 1962a [29]), they do not indicate connotative values (Lakoff, 1973 [151]), they are too vague to allow encoding (Apresjan *et al.*, 1969), they do not rest on any solid theoretical foundation (Pottier, 1965), they do not identify which of the meanings of polysemous words they use (Mel'čuk, 1988 [172]), etc.
> (Béjoint, 1994 [176])

Given all these objections, is there any value in dictionary definitions, or should they be rejected entirely as semantic representations? It is not enough to point out that dictionary definitions have proven their functionality over the past few thousand years: *"Perhaps lexicographers are complacent because their product 'works.' But it is legitimate to ask in what way it works except that dictionaries sell."* (Weinreich, 1962b [26]). Notice that many of the objections above are raised in the course of introducing a semantic theory: Weinreich (Descriptive Semantics), Apresjan & Mel'čuk (Meaning ⇔ Text Theory), Pottier (Interpretative Semantics).

Not all of the criticism of lexical definitions in dictionaries applies to SIM*u*LLDA. For instance, Mel'čuk's point about polysemy, though very valid indeed and related to some problems mentioned before (see 3.3.1), does not apply since they are naturally resolved by the system. And the same holds for Wierzbicka's point about circularity. However, that still leaves a lot of harsh criticism. Let me try to review these various objections in order to save dictionary definitions as a methodologically sound starting point for a MLLD.

In Defence of Dictionary Definitions

Many of the objections claim that dictionary definitions are lacking something or another, such as the points made by Lakoff and McCawley. Lakoff (1973) claims that dictionaries lack connotational information: if we say *"Sarah is a regular spinster"*, we do not say that Sarah actually falls under the definition of spinster (a woman still unmarried beyond the usual age of marrying), but that she behaves like if she were one. Likewise for *"John is almost a fish"*. He calls words like regular and almost 'hedges', and claims that they pose the following problem for dictionary definitions:

> There is no way that, say, a non-speaker of English could, looking at [the] dictionary definitions, figure out what [*she is a regular spinster*] means... The reason is that dictionaries don't usually include connotational information. (Lakoff, 1973 [151])

McCawley claims that *"Dictionaries at present don't give the foreign user any way of knowing that* rather *and* very *cannot be used the same way"* (McCawley, 1973 [168]), because dictionary definitions do not contain rich enough

syntactic information to produce correct sentences. Hence they *"contribute to the prevalence of oriental-sounding English and occidental-sounding Japanese"* (ibid.)

Of course both the grammatical role and the connotation of a word are very relevant. However, they are not of (central) concern to SIMuLLDA, since neither of them is related to the interlingual meanings: connotation (in the intended sense) belongs to the concept; and concepts are, as observed earlier, related to but not identical to interlingual meanings[44]. And information about grammatical usage concerns the word-form, and not the interlingual meanings. Both types of information could be added externally to the system if desired, at the appropriate places. However, the concern in SIMuLLDA is the validity of the dictionary definition in terms of genus et differentiae.

Criticism such as that by Apresjan *et al.* and Pottier is more serious: although the definitions given by dictionaries seem intuitively clear and helpful, on closer inspection they contain many vague descriptions. This criticism involves a pragmatic question: are dictionary definitions indeed too vague to allow encoding? In chapter 4 this question will be addressed, since there an actual encoding of dictionary definitions will be given. The conclusion will be that despite a great number of problems, it is in fact possible to encode dictionary defintions despite some vaguenesses.

Background Knowledge

The definition of a word in terms of genus et differentiae specificae defines one meaning in terms of another. But as we observed in section 3.3.1, this can lead to a serious problem: in a significant number of cases the meaning that is used in the definition is not present as such in the dictionary, as in the case of porthole and window. And this not because of a gap in the dictionary that could easily be filled, but because the idea that words have a fixed, listable number of senses is not entirely appropriate. This effectively means that dictionary definitions, at least in some cases, do not and cannot form a completely self-contained system.

Still, few people will have problems with the LDOCE definition of porthole. And there is a simple reason for that: most people will simply know what a window is, so even if the word window would not be present in the dictionary, the definition would still suffice. And this, I claim, is a fundamental assumption of dictionary definitions: dictionary definitions only 'work' if you have sufficient background knowledge.

A nice example, also discussed by (Hanks, 2000 [9]), is the word googly. The definitions in table 3.11 in principle define what a chinaman is; but,

[44]Some form of connotation, namely that which is indicated by (usage) labels, will be included in SIMuLLDA as explained in the next chapter.

unless you have sufficient background knowledge, these definitions do not help much. You have to know already that they are cricket terms, and have knowledge about what a bowl is, and what the leg side and the off side of a batter are.

chinaman a left-hander's googly

googly a bowl that starts as a leg break but turns into an off-break

off-break a pitch which, after bouncing, breaks into the batter's body from his off side

leg break a pitch that breaks into a batter's body off the bounce, from the batter's leg side

Table 3.11: googly and related words

A result of the dependence of dictionary definitions on background knowledge is also that a dictionary does not provide the proper information for learning a language from scratch: you do not learn Hungarian by carefully considering a Hungarian dictionary.

The SIMuLLDA system does not pretend to provide a solution to this limitation of the functionality of dictionaries: SIMuLLDA tries mainly to provide a system in which definitions as they are currently given in dictionaries can be modelled in such a way that bilingual dictionaries can be generated. So the fact that dictionary definitions are not completely self-contained is inherited by the definitions in terms of definitional attributes.

3.5 Conclusion to Chapter 3

In this chapter, I have specified the interpretation of the basic elements of the SIMuLLDA system: words, languages, interlingual meanings and definitional attributes. The complete set-up including the additional structure for these basic elements illustrated in 3.6. Let me summarize the relevant features of the different parts.

The interlingual lattice consists of an FCA concept lattice, with interlingual meanings and definitional attributes as its formal objects and formal attributes. The languages consist of lexemes, which in turn consist of arrays of word-forms, represented by a citation-form. The languages and the interlingual lattice are connected in that every citation-form of every language is related to one or more of the interlingual meanings, and that every definitional attribute is related to a word-form in every language.
The word-forms consist of a number of pre-word-forms (consisting in turn of a spelling-cum-syllabification and a pronunciation) plus a wordclass plus

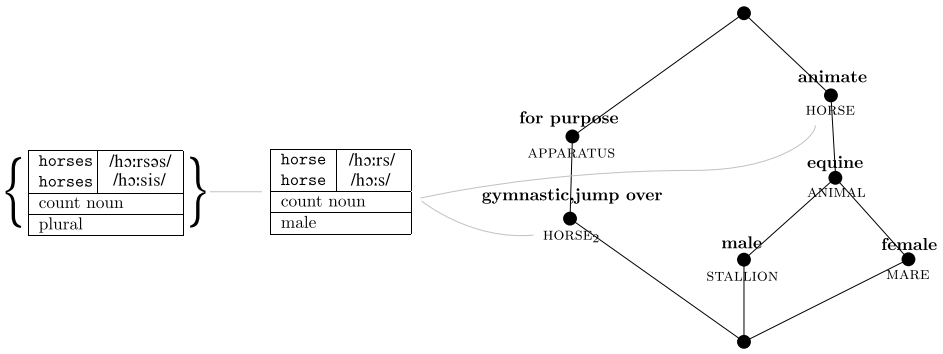


Figure 3.6: Partial Setup of SIMuLLDA

a gender (if applicable). The languages are logically little more than arbitrary sets of lexemes: it is up to the lexicographers to decide both which languages (as opposed to dialects) should be distinguished, and which word-forms should be considered part of the language.

The interlingual meanings are best characterised by what they are not: they are not language-dependent, not denotational in nature, not perceptual in nature, not prototypical in nature, and they can be shared by various languages. Interlingual meanings are basically little more than sets of definitional attributes. Definitional attributes are not Katz & Fodor style semantic markers: they are claimed to be neither psychologically real, nor innate, nor denotational in nature, nor universal, nor interestingly indivisible.

A word (or lexeme) is homonymous/polysemousin SIM*u*LLDA iff it is related to more than one interlingual meaning. Although in principle a sound and standard approach, this leads to a serious problem, because it assumes incorrectly that words have enumerable lists of senses. In reality, the meanings of polysemous terms are overlapping and intertwined. A result of this is that the genus term in a definition is often (especially in cases of regular polysemy) a word-sense that is not as such one of the listed meanings of that word. For SIM*u*LLDA, and in fact any other formal approach, this presents a serious problem and situations like these should be avoided.

With this core set-up of SIM*u*LLDA, translations for words can be found: find the interlingual meaning connected to the citation-form of the desired word-form, and see if there is also a citation-form related to it in the target language. If so, this will be the translation of the word. More interestingly, a translation can also be construed if no lexicalisation of the interlingual meaning exists in the target language (i.e. if there is a lexical gap): find the first superconcept of the smallest common concept of the interlingual meaning for which a translation is required. Then, find the definitional

surplus of the smallest common concept w.r.t. this superconcept. The lexicalisation of the superconcept, together with the lexicalisation of the definitional surplus will be the desired translation; it will be an explanatory equivalent and not a translational equivalent. The monolingual definitions of words can be found in the same way, by taking source and target language to be the same.

In chapter 4, the SIM*u*LLDA set-up will be empirically tested: some real dictionary data will be used to test whether this set-up actually produces the desired results.

Chapter 4

Field Testing: Some Actual Dictionary Data

As explained in chapter 1, the purpose of this thesis is a practical one: to construct a system capable of generating bilingual dictionaries for arbitrary pairs of languages present in the MLLD. And as explained in section 2.3, the starting point for the MLLD are the data found in monolingual dictionaries. So the question is whether this practical purpose is met by the SIM*u*LLDA set-up as described in the previous chapters. The only way to really get an answer to this question is to look at the actual implementation of the system. But in the absence of a full implementation, this chapter will provide the next best thing: a field test, in which the system will be tested against some actual dictionary data.

This field test will consist of three parts; in section 4.1, the data from some different languages for the original example of chapter 2 will be considered: the words for horses. After that, in sections 4.2 and 4.3, two other lexical fields will be discussed in more detail: the words from a number of languages for bodies of water, and the words for the sails on a ship. The data for the second part of the field test (bodies of water) are given in appendix A.

4.1 Horses and the Like

In chapter 2, we introduced the basic machinery for Formal Concept Analysis using words for different kinds of horses as an example. The claim there was that the English words for horses in the LDOCE dictionary were defined in the same fashion as their translational synonyms in the respective dictionaries of these languages. In this section, I will discuss for two of the languages in table 2.6 (Italian and Russian) whether or not this claim holds, with the help of actual dictionary data. The relevant entries for this question are given in table 4.1. The left column displays the definitions

from the Garzanti Italian dictionary, and the Ozhekov Russian dictionary. The right column gives English glosses of these definitions.

cavalla [-là-] *s.f.* femmina del cavallo; giumenta	**cavalla** *f.* female of the horse; mare
giumenta [-mén-] *s.f.* cavalla da sella	**giumenta** *f.* mare for riding
puledro [-lé-] *s.m.* giovane cavallo, assino o mulo, non ancora domato	**puledro** *m* young horse, donkey or mule, not yet domesticated
stallone [-ló-] *s.m.* cavallo maschio destinato alla riproduzione	**stallone** *m* male horse meant for reproduction
жеребе́ц, -бца́ *м.* Самец лошади, достигший половой зрелости	**zherebets** *m* male of the horse, having reached sexual maturity
жеребёнок, -нка *мн.* -бя́та, -бя́т, *м.* Детёныш лошади, а также нек-рых других копытных (ослицы, лосихи, верблюдицы)	**zherebyonok** *m* infant of the horse, and some other ungulates (donkeys, moose, camels)
кобы́ла, -ы *ж.* **1** Самка лошади	**kobyla** *f* female of the horse

Table 4.1: Russian and Italian Words for Horses

Broadly viewed, these definitions do indeed very much resemble their definitions in LDOCE (see table 2.4 on page 36). However, there are some differences that should be discussed. The first is that contrary to table 2.4, Garzanti has no entry for the word puledra (filly). This is not a peculiarity of the Garzanti dictionary: from most Italian dictionaries the word puledra is absent, because it is the regular female version of puledro. Regular derivations have a predictable meaning, and need not be explicitly listed in a monolingual dictionary. If they are listed at all, they are mostly listed as *run-ons*: implicit lexical entries without a definition (more on this in section 5.1.2). But what is a regular derivation in one language does not have to be one in another. Therefore, regular derivations are more commonly present in bilingual dictionaries. For instance, puledra is listed as the translation for filly in the Oxford-Ragazzini. Since SIM*u*LLDA is a multilingual system, regular derivation will need to be added in such cases. Although section 5.1.2 will present a more structural way of adding such a derivation, I will here simply assume that there is a definition for puledra, reading *"giovana cavalla"*.

The second thing is that in all languages, the word for foals applies not only to young horses, but also to the young of other animals. This means that in a way, foal might strictly speaking not be a hyponym of horse (we will discuss a similar case in the next section in more detail). Also, the dictionaries do not agree on which other ungulates the term can relate to. It is hard to establish whether or not this is truly a cross-linguistic variation.

Notice also that the definitions of the translational synonyms of stallion do not contain reference to such a broader applicability.

Thirdly, there is an apparent mismatch between жеребе́ц and stallion on the one hand and stallone on the other: whereas Ozhekov and LDOCE define the requirement that stallions have to be 'adult' (expressed in different ways, which I will get back to), Garzanti makes no such claim. As a result, a direct analysis of these data in SIM*u*LLDA would lead to a situation in which жеребе́ц and stallion have a different set of definitional attributes than stallone has. This would imply that stallion and stallone would not be translational synonyms. More concretely, **STALLION** would have **adult** as a definitional surplus over **STALLONE**, so stallone would be a hyperonym of stallion. But it is dubious whether this would be appropriate: it is more likely that the difference between the definitions of stallion and stallone is not so much a result of the difference between the Italian and the English words, but a result of a different choice of the editors of LDOCE and Garzanti respectively. This suggestion is supported by the fact that we find similar differences between the various English dictionaries on this point, as illustrated in table 4.2.

stallion /ˈstæljən/ *n* a fully-grown male horse kept for breeding – compare MARE (LDOCE)

stallion /ˈstæliən/*n* a fully grown horse, especially one used for breeding (OALD)

stallion, *stal*ˈ*yən*, (*obs*) *n* uncastrated male horse, esp. one kept for breeding (Chambers)

stal·lion (stalˈyən) *n* An uncastrated male horse (Collier)

stal·lion \ˈstalyən\ *n* -s [ME *stalion, stalon* fr. MF *estalon* of Gmc origin; akin to OHG *stall* stall - more at STALL] **1 a:** a male horse not castrated: a male horse kept for breeding; *also* : the male of any equine mammal ⟨a zebra ~⟩ **b:** the male of any of various other animals (as dogs, sheep) kept for or considered in respect to its worth as a stud (Webster)

Table 4.2: English Definitions for *stallion*

Whereas LDOCE and OALD define stallions as fully-grown animals, Collier, Chambers and Webster have no such predicament. The same holds for the question whether or not a stallion has to be uncastrated, and whether or not they have to be kept for breeding.

There seem to be two different sources for these differences of opinion: competing words, and polysemy. To start with the case of competing words: since it is possible to refer to young male horses with the word colt, the word stallion will be oriented towards the other male horses, i.e. the adult ones. But the question is, whether the word colt is hence an antonym of stallion, or simply a hyponym in the fact that the word stallion is pragmatically less applicable to young male horses. It would be a hyponym if using the word stallion for young male horses would merely be less informative than could be, and an antonym of such a use would actually be

incorrect. And the difference between these two situations can be rather subtle, which is in a way illustrated by the fact that LDOCE concludes antonymy, and Collier hyponymy. My personal view would side with Collier and OALD on this matter, especially since colt is a rather technical, not widely known term. But as mentioned before, it is up to lexicographers to decide on the matter.

For the definitional attribute **not_castrated** the same holds as for **adult**, except that the competing word is one not mentioned thus far:

gelding /ˈgeldɪŋ/ *n* an animal, usu. a horse, that has been gelded (LDOCE)

Also here, castrated horses are more commonly denoted by the word gelding, making the hyperonym stallion more prominently associate with specifically non-castrated male horses. Whichever the most appropriate analysis for these two definitional attributes, I would not expect this to lead to cross-linguistic differences.

As said, there is an additional problem with the definitional attribute **adult**: although both LDOCE and Ozhekov somehow indicate that stallions have to be adult, they define the attribute in a different way: Ozhekov claims that they are animals *достигший половой зрелости* (having reached sexual maturity), whereas LDOCE mentions they have to be *fully-grown*. Here even more than between English and Italian, this seems to be an editorial rather than a linguistic matter. In principle there is not even a need to resolve this difference in SIM*u*LLDA: it is possible to have the strings `fully-grown` and достигший половой зрелости as the English and Russian lexicalisation of the definitional attribute **adult**; but it would, at least conceptually, not be attractive to have lexicalisations with a different meaning for the same definitional attribute[1]. So also for the alignment of stallion and жеребе́ц, a choice between the two ways of defining 'adult' has to be chosen.

The suggestion that stallions have to be *"kept for breeding"*, might be instigated, following the definition in Webster, by the fact that stallion is ambiguous between the male horse, and the male of various animals kept for breeding. So the 'esp.' in the Chambers definition would be very appropriate.

In all theses cases, it holds that *someone* has to decide whether or not adult, uncastrated and breeding are part of the definition of stallion and/or its translational synonyms in other languages. So the selection of definitional attributes in SIM*u*LLDA is clearly problematic. But it is only slightly more problematic than the formulation of definition in current monolingual dictionaries. The additional difficulty is that one has to check thoroughly whether or not there are cross-linguistic differences in the defini-

[1]Unless animals are fully-grown exactly when they have reached sexual maturity.

tional fine-print. And this checking should be done by people with expertise on the issue: lexicographers.

So it seems that the data found in actual dictionaries are reasonably in correspondence with the example in section 2.3.1, although the precise analysis should be decided upon by lexicographers. As this example has shown, the actual analysis of the data is a tedious project. It should therefore be repeated that the set-up of SIM*u*LLDA is not to incorporate any existing dictionaries, but to provide a framework for an MLLD. Adding a certain language to the MLLD should take about as long as the compilation of a normal dictionary takes.

There are, of course, many more terms for kinds of horses besides terms for young, old, male, female, and castrated horses. English has the word roan for horses having a roan coat, pinto for a kind of spotted pony, and also words like bangtail, bay, charger, chestnut, eohippus, gee-gee, hack, hackney, jade, mesohippus, mount, nag, pacer, palomino, plug, pony, polo pony, poster, protohippus, riding horse, saddle horse, sorrel, stablemate, stalking-horse, steed, and wild horse for other kinds of horses. It would be interesting to see the problems concerning these other words for horses in SIM*u*LLDA. However, words for animals are often considered 'unfair' for such a test, since they are supposed to be more hierarchic in nature than other words. Hence we will leave these other words for horses for what they are, and turn to a different lexical field, namely all terms for bodies of water.

4.2 Bodies of Water

It has often been claimed that the only concepts that actually do have a hierarchical ordering are what are called *natural kind terms*: terms for things that exist naturally – in contrast with artefact (Quine, 1969). And that hence flora and fauna terms, like the words for horses, are the exceptional cases in which hierarchical definitions actually work. There is a number of reasons why this criticism is not entirely valid. Firstly, the very existence of natural kind terms is problematic. Secondly it is not entirely clear how and why natural kind terms would be different from other concepts. And thirdly, not all terms for horses are natural kind terms: many of the words for horses mentioned above, like riding horse, work horse, and stablemate, are defined functionally. Nevertheless, the discussion of terms for horses would be an unfortunate choice because it cannot be excluded that the semantic field to which the terms apply has a special sort of structure due to the taxonomic form of the biological theory from which they are derived.

Therefore, this section will consider a different lexical field: all words for 'bodies of water', such as rivers, seas, lakes, waterfalls, and lagoons, amongst many others. Three points motivate this choice. The most im-

portant one is that the English words river and stream, and their French (non-) counterparts fleuve and rivière are often quoted as a difficult case for translation:

> Any concept can be refined into more specialized subtypes by making more detailed distinctions. Since different cultures may be sensitive to different features, their languages may have words that have to be translated into other languages either by rough approximations or by clumsy paraphrases. In English, for example, size is the feature that distinguishes river from stream; in French, a fleuve is a river that flows into the sea, and a rivière is either a river or a stream that flows into another river. (Sowa, 1993 [246])

The second motivation is that Ganter & Wille take bodies of water as an example to illustrate the FCA system on the cover of their 1996 book, as represented in figure 4.1. They do not give an explanation of the lattice, other than indicate it as *"Ein additives Liniendiagramm des Begriffsverbandes zu einem Wortfelt 'Gewässer'."*[2] (Ganter & Wille, 1996 [76]). Nevertheless, it suggests that FCA might be well fit for correctly modelling the German words involved.

Notice, however, that it is modelled in a way which is rather unlike the SIMULLDA approach: there are no hyperonyms in the context, so all terms come out as hyponyms of Gewässer. An example of a lexical definition that could be derived from this lattice would be: **Pfütze**: *temporäres, binnenländisches, natürliches, stehendes Gewässer* (a temporary, inland, natural, still body of water). Although the formal objects in this lattice are words, they are treated in a partly denotational way.

The third motivation is that words for bodies of water are words that are often considered difficult to define, in the sense that dictionaries often feel a need about to define them partly by means of illustrations. An example is given in figure 4.2, where the illustration for parts of rivers from Garzanti is given.

In appendix A, the definitions of words for bodies of water from a number of different sources are given: the first is the list of all hyponyms of water generated by WordNet 1.6. After that are the dictionary definitions from a number of dictionaries of comparable size: the English words from LDOCE, the German words from Duden (Universalwörterbuch), the Italian words from Garzanti, the Russian words from Ozhekov, the French words from Robert. The Dutch words are taken from a considerably larger dictionary: the GVD. These data will provide the benchmark test for the system.

[2] An additive line diagram of the concept lattice for a lexical field 'Bodies of Water'.

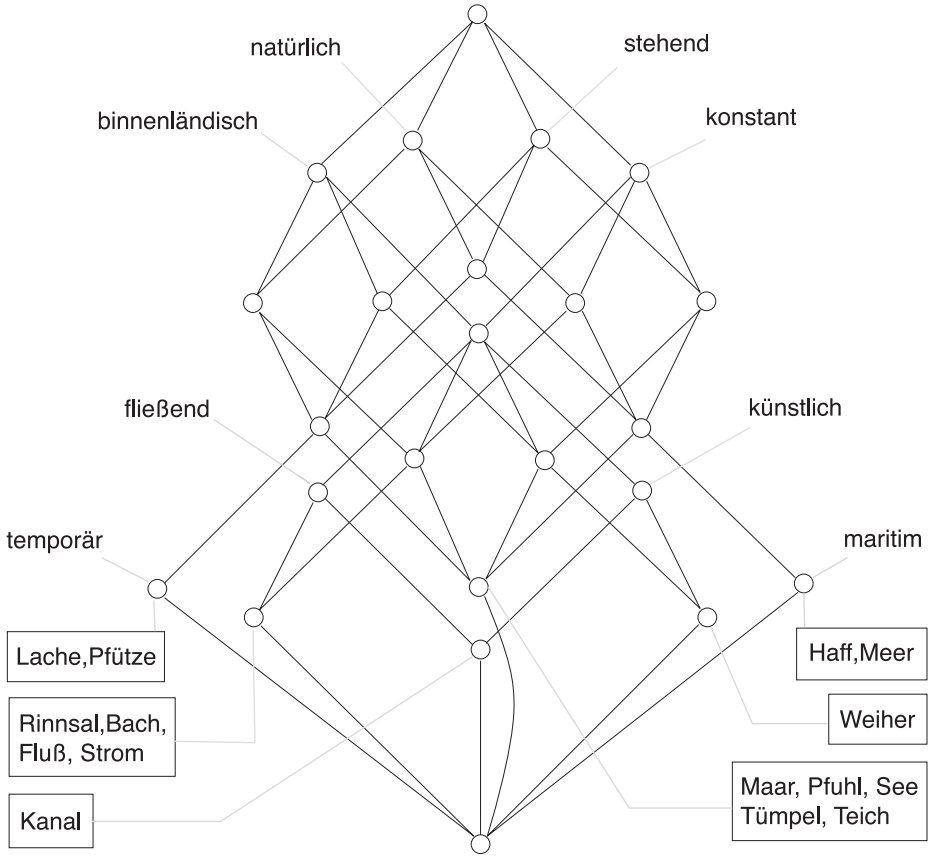


Figure 4.1: Cover Illustration of Ganter & Wille (1996)

The checking of these data will still be rather theoretic, since there is no actual program for testing the data, and the resulting lattices are too large to fruitfully depict; however, the comparison will be described as if the data were fed into an existing bit of software.

The question whether the SIM*u*LLDA system would be able to cope with these data depends on three relatively separate issues: firstly, whether it would be possible to construct concept lattices for the data from the several languages, secondly, whether these lattices would have enough overlap to get a truly interlingual lattice, and thirdly whether this lattice would yield the translations needed for a bilingual dictionary. For the first question, we will consider only the English data found in section A.2 of the appendix. To make easy reading, the relevant definitions from section A.2 of the appendix will be repeated in the text. After this monolingual test, the commensurability will be tested on a few likely problematic cases; and after that the resulting translations will be briefly sketched.

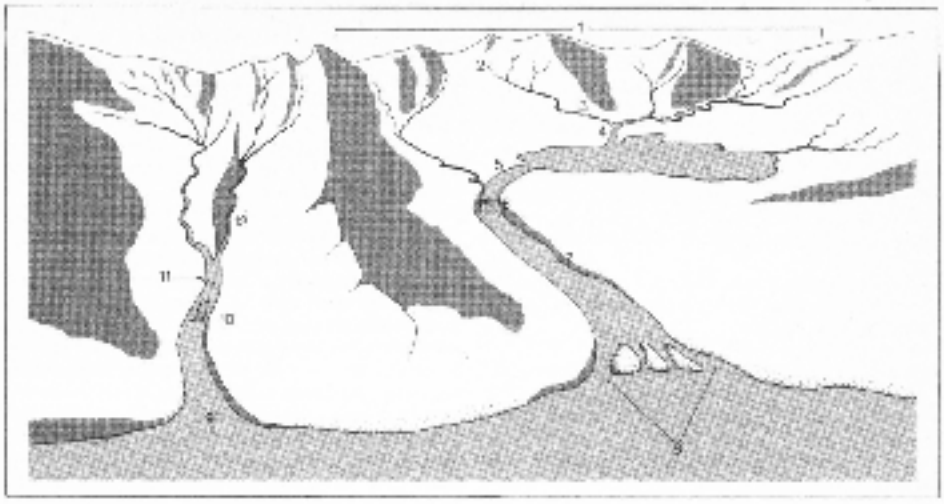


Figure 4.2: Garzanti Illustration for *fiume* (river)

4.2.1 Hierarchy Problems

The first test of the system is to see whether it is possible to construct a SIMULLDA concept lattice on the basis of the definitions in the monolingual dictionary, as given in appendix A.2. Treating these monolingual lexical entries as a lexicographic context should yield a structure that is richer than, but also in many ways similar to, the structure for words for bodies of water in WordNet 1.6, as represented in appendix A.1.

Since all words in appendix A.2 are supposed to be hyponyms of body of water, we can treat body of water as an empty genus term, expressing only a definitional attribute (say: **containing_water**). Within this lexical field, there are a number of lexical entries that can directly be reduced to definitional attributes without much problem, such as the definition of bayou:

> **bayou** /baɪu:/ *n* (esp. in the southeastern US) a body of water with a slow current and many water plants

This definition will reduce to the definitional attributes **with_slow_current** and **with_many_plants** (and **containing_water** of course) in a straightforward way.[3] But there is a large number of lexical entries with some additional issues. In this subsection, the various types of lexical definitions will be looked at, and their possible problems discussed.

[3] The claim is not that this set of attributes is unproblematic, but that the process of reduction is.

Synonyms

There are a number of synonymous words in appendix A.2. Synonyms are not problematic with respect to the interlingua: since synonymous words will refer to the same interlingual meaning, they count as a single entity in terms of structure. But what can be observed is that not all synonyms are defined in the same way. Take the definitions of briny and deep:

briny /ˈbraɪni/ *n* [*the* + S] *lit or humor* the sea
deep[3] *n poet* the sea

The word briny, as well as deep, is defined as a synonym of *sea*. On top of the meaning of sea, it is assigned a register label: *lit or humor*. Since labels are not treated at the level of the interlingua (see section 3.3.4), this has no effect on the fact that briny will refer to the same interlingual meaning as sea and deep. In functional terms: $\text{wfs}_{Eng}(\text{SEA}) = \{\text{sea, briny, deep}\}$. For the same reason, we get: $\text{wfs}_{Eng}(\text{LAKE}) = \{\text{lake, loch, mere}\}$.

But the words brook, rill, and runnel are all individually defined in terms of stream, and not in terms of each other. Nevertheless, they are synonymous, since they all have exactly the same definition:

brook[1] /brʊk/ *n* a small stream
rill /rɪl/ *n poet* a small stream
runnel /ˈrʌnl/ *n esp. lit* a small stream

So even though these words are not explicitly defined as synonyms, they still refer to the same meaning, since a meaning is no more than the set of definitional attributes defining it: $\text{wfs}_{Eng}(\text{BROOK}) = \{\text{brook, runnel, rill}\}$.

Both these kinds of synonymous definitions can be dealt with correctly in SIM*u*LLDA, and will be treated analogously. Conversely, starting from the concept lattice, it is also possible to generate both kinds of definitions for synonyms. The only possible problem is the following: the system cannot distinguish between the two kinds of definitions. If we want to generate a definition for a given word, the system has no way of choosing which kind of definition would be most appropriate.

It is simple to come up with a solution: both briny and deep are uncommon words for a sea, which is indicated by their having a label. These labels can be used to choose between the two alternative kinds of definitions: briny can be defined in terms of sea since it is a non-neutral synonym of sea. Conversely, sea cannot be defined in terms of briny for two reasons: sea is not a labelled word (in which case we could forbid definition in terms of a synonym), and briny is a labelled word (in which case we could forbid it to operate as a synonymous definiens). This mechanism has two advantages: it avoids circular definitions (since in this way, it can never happen that briny is defined in terms of deep, and deep in terms of briny), and it does so in a consistent way.

But the problem is that the result of this mechanism does not correspond to the definitions in LDOCE: rill has the same label as deep, and brook is as unlabeled as sea, yet rill is not defined in terms of brook. So in order to stay true to the definitions in LDOCE, a different system would be necessary. It is the same problem as we saw earlier with the definition of colt: what is the most appropriate definition, *a young male horse*, *a young stallion*, or *a male foal*? It is an open question what the appropriate solution for these questions is.

Notice that all this does not largely affect the effectiveness of the SIM*u*LLDA set-up: the problem is not that the interlingual concept lattice provides incorrect information, it is only a matter of what is the best way to exploit this information. Furthermore, the problem only arises if we want to generate an actual bilingual definition from the system. So the problem is a specific product of the aim of SIM*u*LLDA: the fact that (Euro)WordNet does not have to deal with this problem is that WordNet never tries to generate actual definitions.

Dependent Attributes

Many of the definitions in appendix A.2 contain dependent attributes, all indicating the size of the body of water: **narrow**, **broad**, **small**, and **large**. The following word-forms are defined in terms of one of these attributes:

narrow: channel, creek$_2$, crossing, firth, fjord, inlet, strait, tideway
large: cataract, gulf, lake, sea
small: brook, cove, creek$_2$, crossing, pool, tarn, waterhole

As observed in section 3.4.3, dependent attributes cannot be taken independently from their genus term. Compare for instance the attribute **narrow** in the following two definitions:

> **strait**[1] /streɪt/ also **straits** *pl.* – *n* (*often cap. as part of a name*) a narrow passage of water between two areas of land, usu. connecting two seas
> **creek** /kriːk/ *n* **2** *AmE* a small narrow stream

A stream that would be as broad as a strait would definitely not be *a narrow stream*. As explained (see page 102), this is resolved by restricting the attributes explicitly to their genus term. So channel will not simply have the attribute **small**, but the more elaborate **small <u>rest</u>(sea passage)**, creek will have **small <u>rest</u>(body of water)**, etc. All the basic cases of dependent attributes can be correctly dealt with in this fashion.

But there are some more complex cases of dependent attributes. Apart from the word-forms that are directly defined in terms of **narrow** and **small**, there are two other word-forms that are more complexly defined:

pond /pɒnd||pɑːnd/ *n* an area of still water smaller than a lake, esp. one that has been artificially made

stream[1] /striːm/ *n* **1** *poet* a natural flow of water moving across country between banks, narrower than a river

The complex attributes are **smaller than a lake**, and **narrower than a river**. These are hard to define in terms of restricted attributes, since they do not indicate their size relative to their genus term, but to the size of something else. But on the other hand, there is no need to define them in terms of restricted attributes: the turn of phrase *smaller than a lake* is not dependent at all. It can simply be taken as a atomic definitional attribute, disregarding the internal structure of the attribute[4].

Another interesting case concerns the definitions of sound and rivulet:

sound[5] *n* **1** a fairly broad stretch of water, mostly surrounded by coast

rivulet /ˈrɪvjʊ̈ˑləˈlət/ *n lit* a very small stream

The interesting attributes here are the scaled variant of broad and small: **fairly broad**, and **very small** respectively. Given the fact that BROOK has **small <u>rest</u>(stream)** as an attribute, the attribute **very small** on rivulet somehow indicates that a rivulet is *even smaller than a brook*. If **very small** were simply treated as an elementary attribute, this meaning would not become apparent. However, this problem only shows up because in the SIM*u*LLDA set-up, the definitions of brook and rivulet are put more closely together than in their paper counterparts. The idea behind SIM*u*LLDA was not to improve the existing dictionary definitions, but to put them to better use. So although the richer structure of SIM*u*LLDA might lead lexicographers to give a different definition in such cases, these attributes will get special treatment, and **fairly broad** is simply taken as a dependent attribute, with no relation to **broad**. The word sound will have **fairly broad <u>rest</u>(stretch of sea)** as attribute.

Coordinated Genus Terms and Attributes

According to Vossen & Copestake, almost 20% of all noun senses in LDOCE have a coordinated genus term (Vossen & Copestake, 1993 [266]). That means, they are not defined with a simple genus term, but their genus terms contains an operator like the Boolean *and* or *or*. In the appendix, their number is much lower than 20%, and none with any other construction than *or*. But there are 5 coordinated definitions: estuary, lough, tarn, tributary, and water.

Since the SIM*u*LLDA analysis depends on the semantic definition of the genus term, and such disjunctive definitions do not have a unique genus

[4] Although this becomes rather strange if you consider the Dutch definition of sloot for instance, which is defined as *an artificial water, narrower than a* gracht, *but wider than a* greppel.

term, the system cannot deal with such definitions. However, one could argue that such disjunctive definitions are abbreviated polysemies: if a word is defined as *an (a or b) having c*, it can have either of two meanings: *an a having c* or *a b having c*. This can be nicely illustrated by the definition of borer in LDOCE[5]:

bor·er /ˈbɔːrə^{r}/ *n* a person, tool, or insect that makes round holes

This definition can be seen as actually giving three distinct meanings to the word: a borer can either be a tool to make holes with, a person whose job is making holes, or an insect that gathers food by drilling holes. The translation in Dutch for these three meanings would be different: boor for the tool, boorder for the person, and borend insekt for the insect (according to VDEN). The reason why these different meanings can be piled together in a single definition is that they on top of being homographs, they are named borer for the same reason: because they bore holes. But in a system like SIM*u*LLDA, these three distinct meanings should be treated as different interlingual meanings. So the solution is to take such abbreviated polysemies apart, after which the coordinated genus term simply disappears.

In the case sketched above, the coordinated genus term is more or less an abbreviated way of giving polysemous meanings. Though this is a possible cause for having a coordinated genus term, more commonly the reason is a different one. Consider the following three definitions:

tarn /tɑː||tɑːrn/ *n* (*often cap. as part of a name*) a small mountain lake or pool, esp. in the north of England

lake[1] /leɪk/ *n* **1** a large area of water, esp. non-salty water, surrounded by land

pool[1] /puːl/ *n* **1** a small area of still water in a hollow place, usu. naturally formed

The genus term of tarn is *lake or pool*[6]. Lakes and pools are not unrelated things, like persons, insects and tools are. In fact, they are *"a large area of water"*, and *"a small area of water"* respectively. This suggests that a tarn can be an area of water of any size, but that for some reason the lexicographers preferred this disjunctive construction over taking a more general word like area of water as hyperonym. A similar situation can be sketched for more or less all the coordinated genus terms in A.2. Given the difference between SIM*u*LLDA and paper dictionaries, this might be a reason to prefer a more general genus term for SIM*u*LLDA. So in the SIM*u*LLDA analysis, it might be best to interpret the definition of tarn as *"a small area of water in the mountains"*. However, there is nothing compulsory about this choice; there are simply two ways of reanalysing coordinated genus terms in SIM*u*LLDA:

[5]Example taken from Vossen & Copestake (1993 [266])

[6]Disregarding the *mountain* bit for clarity here. Notice also that the interpretation of *mountain* in the definition is somewhat unclear: it is presumably distributing over *lake* and *pool*, but although a mountain lake is a normal term, mountain pool is much less so.

either as abbreviated polysemies, or as implicitly giving a more general genus term. It is up to the lexicographer to choice which method is the most appropriate in each case.

More common than coordinated genus terms in appendix A.2 are the definitions with a disjunction in their differentiae specificae. There are twelve such 'disjoint attributes', in the definitions of canal^{a}, canal^{b}, creek, fjord, inlet (2x), lagoon, river, sea^{2b}, tributary, waterfall, and waterway. Disjunctions in definitional attributes do not (automatically) pose a problem: definitional attributes are lexicalised as free text items, and there is nothing that stops these free text items from containing disjunctive constructions. It is of course possible to choose for the same strategy as with disjoint genus terms, and have the disjunction lead to polysemy. In fact, the LDOCE definition of sea contains a disjunctive attribute, which is explicitly intended to be read polysemously:

> **sea** *n* **1** the great body of salty water that covers much of the Earth's surface; ocean **2** a large body of salty water smaller than an ocean, either **a** part of the ocean **b** a body of water (mostly) enclosed by land

But in most cases, creating polysemy for disjunctive attributes will not be a desirable or necessary solution.

There is an additional consideration here: attributes are weakened by disjunctions. So the attribute expressed by *"that flows into a larger stream or river"* (for tributary) is less strong an attribute than *"that flows into a river"* would be. When need arises, such entailments can be modelled with partial order on attributes (see section 2.4.7); however, there are not such cases in appendix A.2, so we do not need to worry about partial orders here.

Meronyms

There is an important group of words in appendix A.2 that are not defined in terms of genus et differentiae: the meronyms. Meronyms are not defined as a special kind of something else, but as part of something else. There are 7 such definitions: lough and loch are indicated as *part of a sea*; pool, channel, rapids, and estuary are defined as *part of a river*, and the wordform bay is defined as *part of a sea or large lake*.

Given the dependence of the SIM*u*LLDA decomposition on the definition of the genus term, these definitions pose a problem. To solve this, there are basically two possibilities: either the system has to be extended to directly deal with meronyms as special kinds of definitions, or the meronymous definitions have to be reanalysed as special kinds of genus et differentiae definitions. Most lexical database systems, such as WordNet, opt for the first option. In WordNet for instance, there is a special relation between synsets for meronymy.

Defining a special relation for meronymy suggests that meronymy provides a special way of defining words/concepts. And that is not (entirely) correct. To show this, let me first make a distinction between defining the whole as having the parts, or defining the part as being part of the whole. To give an example: if we define a meronymous relation between wing and airplane, this can be taken in two ways:

1. a characterisation of airplanes: having wings is an important feature of airplanes

2. a definition of what wings are: wings are the things you find on airplanes

Confusing these can lead to the kind of unclarity mentioned by Soergel in his review of *WordNet* (Miller, 1998): *"'airplane has-part wings' really means 'an individual object 1 which belongs to the class airplane has-part individual object 2 which belongs to the class wing. While it is true that 'bird has-part wing', it is by no means the same wing or even the same type of wing."* (Soergel, 1998).

In both of the cases described above, meronymy boils down to a kind of differentia specifica. For the first interpretation of meronymous definitions this is very straightforward: having wings is simply an attribute of airplanes. And for many such meronymous relations, it is questionable whether they are *definitional* attributes: that cars have steering-wheels is more encyclopaedic knowledge about cars.

Also in the second sense, there is little need for a special meronymous relation: since there are, or might be, other words/concepts that also will be defined as part of an airplane (wings are not *the* things you find on airplanes), it is just a partial characterisation of wings; and that means that it is just a differentia specifica like all other. Saying that wings are parts of airplanes is equal to saying that being part of an airplane is a differentia specifica of wing[7].

Therefore, the route that is chosen in SIMULLDA is the second option: reinterpret meronymous definitions. Meronymous definitions in dictionaries are always of the second kind, and can be reinterpreted as described above. How this works can be nicely shown using the definition of rapids as an example:

> **rapids** /ˈræpə̯dz/ also **whitewater** *AmE* – *n* [P] part of a river, where the water moves very fast over rocks

If we reanalyse this definition in the way indicated above, then its lacks a genus term: it is not explicitly indicated what kind of objects rapids are.

[7]This does not imply that the WordNet approach is wrong, since there is no formal specification of how to interpret the meronymous relations in WordNet. But it does imply that the importance of the **part_of** relation in WordNet should not be overestimated.

The only thing that is indicated is that rapids have two differentiating properties: on the one hand, they are things where *the water moves very fast over rocks*. And on the other hand they are *part of a river*. In this interpretation, the definition is simply indicating these two definitional attributes, without any genus term. So if we add the uninformative something as genus term, the definition would read: *"something, which is a part of a river, where the water moves very fast over rocks."*

In this analysis, it might be more useful to add a genus term: rapids is a hyponym of bodies of water, so the definition for rapids could be changed to: *"body of water, which is a part of a river..."*. Of course, it would be up to the lexicographer to decide what the appropriate genus term would be; following WordNet (see appendix A.1), it could also be the even more specific waterway.

Vossen & Copestake (1993) mention a larger variety of definitions that are not of the form genus et differentiae, but have a 'complex kernel' (with a complex genus term): not only COMPONENT/WHOLE (part of, the system of), but also TYPE/KIND (a type of, a kind of), MEMBER/GROUP (a member of, a group of), and QUANTITY/MASS (an amount of). Though these kinds of definitions do not appear in the appendix, it is clear that a similar treatment would be appropriate. So also in the case of MEMBER, we do not want a special mereological treatment, but a redescription as an *"entity, which is a member of..."*.

Non-Existing Genus Terms

Problematic for the lattice structure are also those word-forms in section A.2 of the appendix that are defined in terms of a genus term that is itself not defined in the dictionary. In the formal set-up of SIM*u*LLDA, these non-existing terms have to be defined in some way, or they have to be circumvented. There are basically three reasons for the absence of the genus meaning:

1. the meaning may be absent accidentally
2. the meaning may be left out on purpose because it is predictable (e.g. stretch of water)
3. the meaning may be left out because of the polysemy problems discussed in section 3.3.1 (e.g. the desired meaning of window)

Genus meanings that are accidentally absent are not particularly interesting. There are no examples of such cases in the appendix, but it could have happened that the word foal is defined in the dictionary as *a young ungulate*, but that the word ungulate itself is not defined in the dictionary. Such accidental gaps could simply be filled by adding a lexical entry for ungulate.

More interesting are the genus terms that are absent because of predictability. An example can be found in the following definition:

waterway *n* stretch of water, e.g. part of a river, which ships and boats can move on

The genus term of this definition is *stretch of water*, which is not present in the dictionary because its meaning can be predicted by combining the meanings of stretch (of) and water. Similarly, there are no definitions for body of water and flow of water, because they too are predictable multi-word units. Predictable multi-words units are not lexemes (but free phrases, see section 5.2.1), and hence have no place in a dictionary. Given the following definitions, the meaning of multi-word units mentioned above should be clear:

body /ˈbɒdi||ˈbɑːdi/ *n* **3** [C (**of**)] a large amount
stream[1] /striːm/ *n* **2** [(**of**)] something flowing or moving forwards continuously
stretch[2] /stretʃ/ *n* **3** [C (**of**)] a level area or SECTION of land or water

Also in the case of such intentional absences, it would be possible to add these expressions to the dictionary (despite their predictable meaning), since they are needed for the definition of their hyponyms. This would solve the problem, but has two disadvantages. Firstly, it is not an elegant solution to add elements to the dictionary because the system used for formalising the dictionary requires them; the only consideration should be whether or not the items belong to the dictionary, and the system should adept to that, rather than change the data.

The second problem is the following: since such phrasal units are compositionally built up, they can be extended in various ways. And we find such modified phrasal expressions in the appendix: beside the simple *area of water* and *body of water*, there are more complex structures: pool and pond are defined *"an area of still water"* and sea as *"a body of salty water"*. And an area of still water is not a still area of water. That might mean that we should also incorporate body of salty water in the dictionary, making the solution even less attractive.

The best solution to these problems is the following: given the fact that the definition of lake is defined in terms of an area of water, the system should logically contain information about what an area of water is. But because of its predictability, it would be a waste of space to have a term in any dictionary. So the phrasal expression area of water will be present in the system, but it will be labelled to indicate that it should not be included as a entry word in a dictionary (such labelling will be discussed in section 5.4). The case of *area of still water* can be analysed in terms of area of water and a definitional attribute: *where the water is still*. This is (almost) an attribute that is already present in the appendix: pond is defined as a part of a lake *where the water is almost still*.

The third, and most difficult reason for absent genus meanings is polysemy, where the genus meaning is absent because it is not one of the meanings listed at the corresponding word-form, but something 'in between' (see section 3.3.1). As observed earlier, this is an important shortcoming of dictionaries from a theoretical point of view, that can now be illustrated by an actual example.

The word cove is defined as *"a small bay"*, where the relevant definition of bay consists of two parts: *wide opening along a coast* on the one hand, and *part of sea or of a large lake enclosed in a curve of the land* on the other:

> **bay**[1] *n* a wide opening along a coast; part of the sea or of a large lake enclosed in a curve of the land
>
> **cove** *n* a small sheltered opening in the coastline; small bay[1]

These two rather different definitions for bay_1 are suggested to form one single meaning, since they are listed under the same sense-number, separated by a semi-colon. The semi-column indicates conjunction, which usually concerns alternative ways of giving the same definition, An example is the Dutch definition of baai, defined as *"small gulf; sea-arm"*. But this suggestion of unity does not make the two definitions any more alike; and the only way to deal with such a definition is by treating it as consisting of two separate definitions. This double definition of bay_1 is a good example of a container/containee polysemy, as discussed on page 83, leading to polysemy/homonymy. But if we treat bay as polysemous, the intended genus meaning of cove is ipso facto absent: the genus term in cove is the combination of these two meanings. And since there is no way of choosing which of these two meanings the word cove should relate to, the only option (however inelegant) is to treat cove as polysemous as well.

As discussed before, this inelegant solution is due to the fact that this absence of the intended genus term is a serious problem of lexicography. The problem is that implicitly, every dictionary by its very design assumes word-senses to be nicely enumerable, while in fact they are not. Given the fact that this enumerability is such a basic assumption behind lexicography, it is hard to see how any lexical database system could resolve this problem in a completely satisfactory way, as long as it is based on lexicographic data.

Summing Up

In this analysis, I have shown that it is in principle possible to analyse the data in section A.2 of the appendix as a lexicographic context. This analysis is not without problems: for most lexical entries, some additional effort was necessary to get it into the SIMULLDA analysis. But these problems are simply an illustration of the fact that word meaning is a very complicated issue. Only one solution was not entirely satisfactory: the problem of (regular) polysemy and the resulting absence of the necessary genus meaning

for the hyponyms of some such polysemous words. This is a basic problem any MLLD has to deal with.

With the solutions presented in this subsection taken into account, the dictionary entries in appendix A.2 can be viewed as a lexicographic context. The formal objects are the 64 word-meanings that are listed, and these 64 word-meanings are defined in terms of 103 definitional attributes, the English lexicalisation of which is listed in table 4.3.

V- or U-shaped	large rest(body of still water)	sheltered
almost enclosed by land	large rest(stretch of sea)	slowing between banks into a lake, another wider stream, or the sea
almost surrounded by land	large rest(waterfall)	small rest(area of still water)
along a coast	larger than, or curving less than, a bay	small rest(area of water)
artificial	long rest(body of water)	small rest(opening)
at which a road, river, border etc., can be crossed	lower	small rest(stream)
between cliffs or steep slopes	lying in a hollow place in the ground	smaller than a lake
between two areas of land	making it dangerous to boats	smaller than an ocean
branch	mostly surrounded by coast	steep
connecting two areas	moving across country between banks	surrounded by land
connecting two larger bodies of water	narrow rest(area of sea)	that connects with a main one
cut into the ground	narrow rest(body of water)	that covers most of the Earths surface
deep	narrow rest(passage of water)	that flows into a larger stream or river
deeper	narrow rest(sea passage)	through which the tide flows
deepest	narrow rest(stretch of water)	through which water flows
dug in the ground to allow ships or boats to travel through	narrow rest(stream)	underwater
dug in the ground to bring water to boats to travel through	narrower than a river	used for driving the wheel of a watermill
enclosed in a curve of the land	natural	very small rest(stream)
fairly broad rest(stretch of sea)	natural or artificial	where a river flows out
falling straight down over rocks	near a countrys coast	where a river makes a sudden deep drop
from which all the water flows into the same river	not far blow the surface of the water	where animals come to wallow
great rest(body of salty water)	of a stream, river, etc.	where it can be crossed on foot, in a car, etc. without using a bridge
great rest(sea)	of sand	where liquid is stored
high	of sea water	where the water in not very deep
in a coast	over which that country has legal control	where the water is almost still
in a hollow place	part of a river	where the water moves very fast over rocks
in a river	part of a river, harbour or sea passage	where wild animals go to drink
in dry country	part of something larger	which do not belong to any particular country
in the coastline	part of the river	which ships or boats can move on
in which foreigners are not allowed to catch fish	part of the sea	wide
into which the sea enters at high tide	part or mouth of a river	wide rest(opening)
large rest(area of water)	partly enclosed by land	wide rest(stretch of water)
	partly or completely separated from the sea by banks of sand, rock, coral, etc.	wider than a strait
	reaching from the sea, a lake, etc. into the land	with a slow current
	separate	with many water plants

Table 4.3: Definitional Attributes for Appendix A.2

The resulting lexicographic context has a total of 132 formal concepts. Even though this number of formal concepts is relatively low, a lattice with 132 nodes is still much too large to fit readably onto a page. Therefore, no graphical presentation of the lattice will be given here, but it can be found on the web at the web-site of this thesis[8].

4.2.2 Interlingual Alignments

So far the data in appendix A were only considered within a single language, which only brings up the problem of building a lattice out of the monolingual English dictionary data. Much more difficult than this monolingual problem is the problem of aligning the various languages, and have all languages in the appendix relate to the same interlingual lattice. The

[8] `http://maarten.janssenweb.net/simullda`

reason for that is that the definitions from the various languages virtually never coincide completely. Take the following two rather similar definitions as an example:

Wasserstraße von Schiffen befahrbares Gewässer als Verkehrsweg
waterway *n* a stretch of water which ships or boats can move on

However close these definitions are, there are differences in every part of them. The genus term is different: Gewässer corresponds more closely to *body of water* than to *stretch of water* (which would have been Wasserstrecke in German). The expression *'von Schiffen befahrbar'* ('navigable for ships') is different from the English *'which ships and boats can move on'* in that both Schiff and befahren do not have a strict English translation: Schiff is indifferent between boat and ship, and befahren is specifically travelling over water done by *Schiffen*. And then there is the additional *"als Verkehrsweg"* (as highway) which is absent from the English, but is partly entailed by *can move on*.

So given the contentual closeness but their descriptive non-similarity of the definitions, the question of whether or not Wasserstraße and waterway refer to the same interlingual meaning is not a trivial one. It could well be that the differences between the German and the English definition do not reflect a difference in meaning between Wasserstraße and waterway, but are simply different formulations of the same meaning, similar to the differences between the definition of waterway in two monolingual English dictionaries. If that is the case (which I believe it is), then the two definitions have to be changed in such a way that they match up; not because either of the definitions is wrong, but because translational synonyms should be defined analogously. The same difficult kind of decision has to be made for every pair of definitions, and all the answers had better not been given by me, but by trained lexicographers.

Nevertheless, in this subsection I will attempt to give a multilingual analysis of the data in appendix A, though the results are definitely not univocal in any sense. As will become clear by the length of the discussion, there is only room to discuss the English and French alignment of river, fleuve, rivière and *stream* in much detail; all the rest of the cases will only be discussed coarsely, giving an idea of the number of lexical gaps in the appendix.

River and Fleuve

The often cited problem of the mismatch between the English words river and stream, and the French words fleuve and rivière nicely illustrates how difficult interlingual alignment is. The proper meaning of these words is hard to assess, since there are many different analyses around. In the tradition of computational linguistics, the basic assumption is that French and

English have different distinguishing features, which results in a complex mismatch between the two sets of terms. In the continuation of the quote cited on page 116, and with reference to figure 4.3, Sowa suggests that we need an intricate solution for this mismatch:

> Figure [4.3 of this thesis] shows the portion of the type hierarchy that includes the lexical types for [these] words and their subtypes. In translating French to English, 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*. In the reverse direction, river maps to RIVER, which has two subtypes: one is FLEUVE, which maps to fleuve; and the other is BIG-RIVIERE, whose closest approximation in French is the word rivière or the phrase *grande rivière*. (Sowa, 1993 [246])

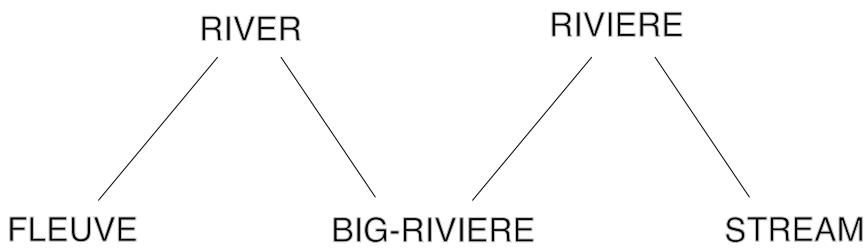


Figure 4.3: RIVER, STREAM, and their French synonyms (Sowa, 1993)

In the domain of interpretative semantics, there is a rather different analysis. Noailly (1996) rejects the idea that French uses the ending of the stream of water as a distinguishing feature, though she does suggest a complex matching between the French terms and their translation in other European languages, such as English[9]:

[9]The analysis by Noailly, which is in the tradition of interpretative semantics (see section 3.4.1), rejects the geographers definitions of fleuve on the basis of it not being experientially based: *"Ce qui est sûr, c'est que nul ne retient, dans l'usage générale, cet essai de catégorisation scientifique proposé par les géographes selon lequel, indépendamment de l'importance du cours d'eau, on devrait appeler* fleuve *ce qui se jette dans la mer, et* rivière *ce qui va dans le fleuve. Cette nomenclature repose en effet sur une connaissance abstraite de la destinée du cours d'eau concerné, et non sur la perception immédiate, et a peu de validité expériencielle."* – What is certain is that nothing in the general use supports the scientific attempt proposed by the geographers according to which, independent of the importance of the stream of water, we should call those a fleuve which end in the sea, and those that end in a fleuve a rivière. That nomenclature depends in fact on an abstract knowledge of the destination of the stream of water in question, and not on the direct perception, and has little experiential value. (Noailly, 1996 [26]) This illustrates nicely that although according to Rastier, sémèmes are not perceptually based, in practice they are nonetheless closely linked to perceptual information.

> À l'intérieur du taxème des cours d'eau, [fleuve] semble présenter le sème spécifique /grandeur/, qui l'oppose à rivière et ruisseau. On peut observer que ce système de représentation tripartite est assez spécifique du français, la majorité des langues européennes se contentant d'un système simplement bipartite sur l'échelle des dimensions, et confondant dans une unique représentation tout cours d'eau d'une relative importance.[10] (Noailly, 1996 [26])

Given the purpose of SIM*u*LLDA, it is the lexicographers perspective on these words that is of primary concern, more than these linguistic perspectives. Since in the SIM*u*LLDA set-up, the interlingua is built from the analysis of the monolingual dictionary data, the proper analysis should in principle follow from the lexical entries of these terms. The English definitions that are relevant to this end, taken from some major English dictionaries, are given in table 4.4, while their French counterparts are given in table 4.5.

river *n* a wide natural stream of water slowing between banks into a lake, into another wider stream, or into the sea (LDOCE)

stream[1] *n* a natural flow of water moving across country between banks, narrower than a river (LDOCE)

river *n* **1 a** a natural stream of water of usually considerable volume (Collegiate)

stream[1] *n* a body of running water (as a river or brook) flowing on the earth; (Collegiate)

river /ˈrɪvə(r)/ *n.* **1 a** a copious natural stream of water flowing in a channel to the sea or a lake etc. (COD)

stream /striːm/ *n.* **1** a flowing body of water, esp. a small river (COD)

Table 4.4: Definitions of river and stream

Let us start the analysis by looking at the French word fleuve: there is no consistent answer in table 4.5 to the question whether ending in sea is a defining property of fleuve. Larousse states that it is, while the Petit Robert says it is not: only as a technical geographer's term does the word fleuve specifically relate to streams ending in sea; as an everyday term, it is simply a major river, where its importance can be due to a number of facts. The definition of the Hachette is somehow in between: it claims that **ends_in_sea** is a definitional attribute of fleuve, but that it is not the only differentia specifica. Only when taken as a geographer's term, **ends_in_sea** is a sufficient property for streams of water to be a fleuve. The SIM*u*LLDA system can

[10] Inside the taxeme of streams of water, fleuve seem to present the specific sème /greatness/, which opposes it to rivière and ruisseau. One can observe that this system of tripartitional representation is rather specific for French, the majority of European languages simply taking a bipartition on the field of dimensions, and grouping all streams of water of a relative importance into a simple representation.

fleuve [flœv] n.m. *-fleuve* XIIe. 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** Big *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 footnotesize) that ends in the sea.
rivière [rivjɛr] n.f. -1138; ruisseau 1105; bas. lat. *riparia* de *ripa* → rive **I**$_{\diamond}$ **1**$_{\diamond}$ Cours d'eau naturel de moyenne importance (Larousse)	**rivière** Natural stream of water of medium importance
fleuve [flœv] n.m. (lat. *fluvius*) Cours d'eau qui aboutit à la mer (Larousse)	**fleuve** Stream of water that ends in the sea
rivière [rivjɛr] n.f. (du lat. *riparius*, qui se trouve sur la rive) Toute espèce de cours d'eau abondant, et particulièrement celui qui se jette dans un fleuve (Petit Robert)	**rivière** Every abundant stream of water, and particularly one that ends in a *fleuve*
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
rivière [rivjɛr] n.f. **I. 1.** Cours d'eau de moyenne importance ▷ *spécial* Cours d'eau qui se jette dans un autre cours d'eau (à la difference du *fleuve*) (Hachette)	**rivière** Stream of water of medium importance *special* Stream of water that ends in another stream of water

Table 4.5: Definitions of rivière and fleuve

of course follow only one of these analyses (like any dictionary system). The definition from Hachette seems the most neutral, so let us opt for that definition: a fleuve is a big stream of water ending in sea.

According to popular belief, the analysis of rivière should follow the same line, except that rivières end in other rivers rather than in sea. But this nice symmetry between the two terms is not found in any of the three dictionaries: the Petit Robert and the Hachette in the case of rivière lists ending in another river as a prototypical feature or 'special' meaning rather than reserved for a geographical term; and the Larousse even makes no mention of ending in another river, but merely defines it as a stream of medium importance. Furthermore, the Petit Robert and the Hachette no longer define a *rivière* as a **big** stream (although the Larousse does label it *abundant*). Still, we will adopt the parallel view, where rivière is defined as a big stream ending in another river for two reasons: firstly, it is the definition in Larousse, and the most prominently present view in much discussion. But more importantly, it is the definition most in correspondence with the data from bilingual dictionaries. The relevant bilingual definitions are given in table 4.6.

fleuve /flœv/ **I** *nm* **1** Geog river (OxHach)
rivière /rivjɛr/ *nf* **1** (cours d'eau) river (OxHach)
ruisseau *pl* ~**x** /rɥiso/ *nf* **1** (cours d'eau) stream, brook (OxHach)
river /ˈrɪvə(r)/ ▷**1644** *n* **1** (flowing into sea) fleuve *m*; (tributary) rivière *f* (OxHach)
stream /striːm/ **I** *n* **1** (small river) ruisseau *f* (OxHach)
rivière [rivjɛːr], *s.f.* **1** (a) river; stream (Harraps)
fleuve [flœːv], *s.m.* (large) river (Harraps)
ruisseau, -eaux [rɥiso], *s.m.* **1** brook, (small) stream, streamlet, rivulet (Harraps)
river ['rivər], *s.* **1** (a) cours *m* d'eau (*entering sea*) fleuve *m*; (*tributary*) rivière (Harraps)
stream [striːm], *s.* **1** (a) cours *m* d'eau; fleuve *m*; rivière *f* (Harraps)

Table 4.6: Bilingual Definitions of river, stream, rivière, and fleuve

Both rivière and fleuve are consistently translated with river in these bilingual dictionaries. And since river is defined as a **large** natural stream, so should fleuve and rivière naturally be. Both dictionaries define fleuve and rivière as translational hyponyms of river, with the opposition between *to the sea* and *tributary* as differentiae, where *tributary* a in turn involves ending in another river.

And finally the word stream: although the word stream is defined in LDOCE as almost the same as a river but smaller, it is not defined as such in the other dictionaries: stream is a more general term, which does not specify size. The remaining words ruisseau and brook (or its synonyms in English) are both differently defined, as **small** streams of water. These considerations lead to the following definitions (given in English lexicalisation for clarity)[11]:

river big natural stream of water
rivière big natural stream of water that ends in another stream of water
fleuve big natural stream of water that ends in the sea
stream natural stream of water
ruisseau small stream
brook small stream

There is a further complication with these definitions: in the multilingual set-up resulting from these definitions, rivière will be a lexical gap in English[12]. But consider the English word tributary. It is more or less defined as a *river that ends in another stream of water*. That would make it simply

[11] Any other interpretation could have been used for the SIMULLDA analysis equally easy.

[12] The phrase '*x* is a lexical gap (in language em Y)' will be used as a shorthand for *there is a lexical gap in language Y for word x*, which in turn means that there is no translational synonym in *Y* for the word *x*.

a lexicalisation of the interlingual meaning RIVIERE. Still, none of the dictionaries consider it a good translation of rivière, though both indicate that river should be translated by rivière *when talking about a tributary*.

The reason why tributary is not a correct translation of rivière might be that a tributary is not defined in terms of **big**: it is a more general term, only indicating the end-point of the stream, and not its size; a meaning for which the French have the word affluent (and possibly the technical use of rivière). If we adopt this analysis, tributary will be a hyperonym of rivière. So we can give the complete picture for these terms in a lattice, as in figure 4.4.

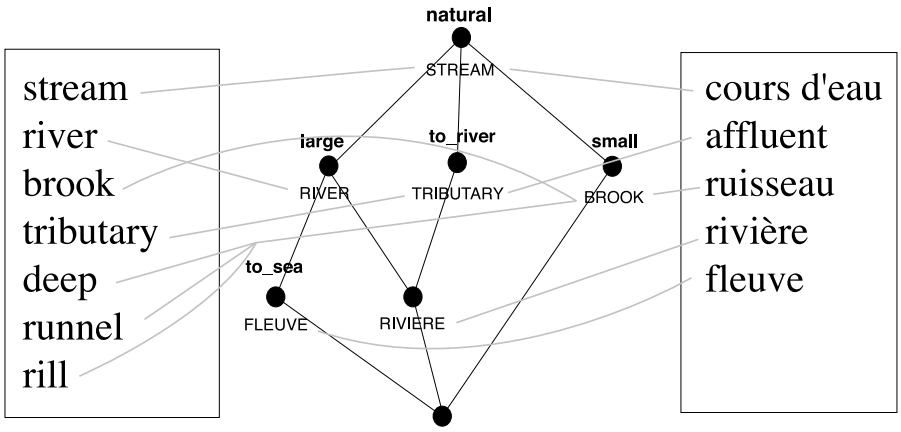


Figure 4.4: Concept Lattice for Streams of Water

It is useful to compare this analysis in SIMuLLDA to the analysis that other systems would give. Take for instance the analysis that EuroWordNet would give, illustrated in figure 4.5[13].

All words that are not lexical gaps are dealt with nicely in this EuroWordNet analysis. For instance, the two synsets {feeder, tributary, affluent} (from English) and {affluent} (from French) both have a relation **eq_synonym** to the Interlingual Item (ILI) RIVER. This means that the English feeder and the French affluent are said to be translational synonyms, like they are in the SIMuLLDA analysis in figure 4.4. In principle, there is also little wrong with the analysis of the lexical gaps: the two French synsets {rivière} and {fleuve} both have a **has_eq_hyperonym** relation to the ILI RIVER, to which the English synset {river} has a **eq_synonym** relation. So by this analysis, both the French rivière and the French fleuve are said to be translational hyponym of the English river.

[13]The French data are speculative, all the others are in correspondence with section A.1 of the appendix and figure 1.5.

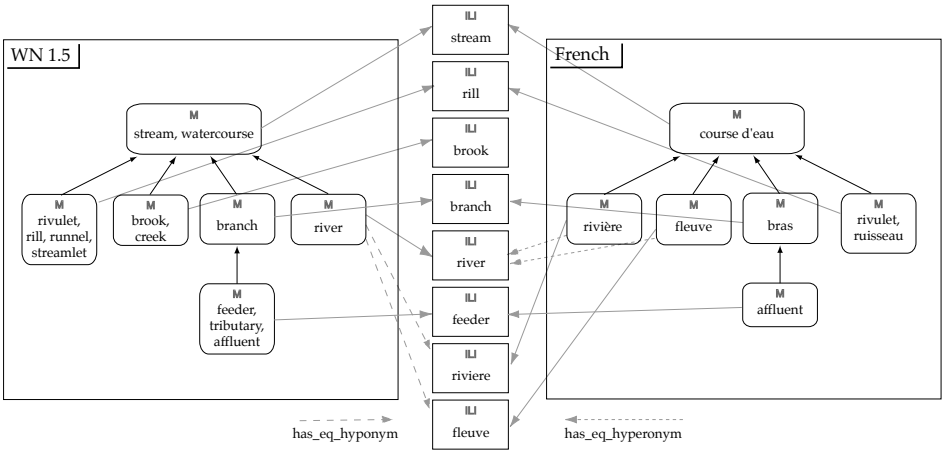


Figure 4.5: EuroWordNet analysis of Streams of Water

But the SIMULLDA analysis has two advantages over the EuroWordNet approach: the hyperonym relation between the English and French terms is much more directly given than in the EuroWordNet representation: there it has to be reconstructed from the nine different relations between the three synsets and the three ILI's. And if more languages with even different attributes would be added, this situation would deteriorate rapidly. But more importantly, in EuroWordNet there is no way of telling in what way fleuve is more specific than river; what its translational differentiae specificae are. There is even no way of telling the two translational hyponyms of river apart.

Comparing the SIMULLDA analysis to the analysis by Sowa in figure 4.3 illustrates three problems of the analysis by Sowa: firstly, Sowa introduces the LU (lexical unit) BIG-RIVIERE. But in neither of the two languages is there a lexicalisation of this LU. So there is ill motivation for introducing such an LU. It helps only very indirectly to link the two words river and rivière. Secondly, Sowa claims that when desired, we can describe fleuve as *river that runs into sea*. But although the figure does specify that FLEUVE is more specific than RIVER, it does not indicate in what way it is more specific. To arrive at the decription *river that runs into sea* we need a definitional attribute, and a lexicalisation of that definitional attribute in English. Both of which are naturally provided by the SIMULLDA analysis. Thirdly, the analysis by Sowa does not conform with the explanation he gives: his claim is that stream is just smaller than river, whereas from figure 4.3, we can conclude that STREAM also has to be defined in term of *flowing to another river*.

There are two additional approaches I would like to mention here, that conceptually are closer to the SIMULLDA set-up than those mentioned before. The first is the NADIA system, as described by Serasset (1994). In the

NADIA set-up, the meaning units are called *acceptions*, and acceptions exist both at the level of the language, as at the level of the interlingua. The representation for some of the words discussed here is given in figure 4.6.

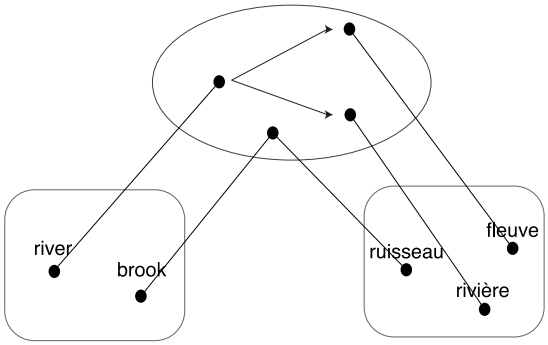


Figure 4.6: NADIA analysis of river, rivière, and fleuve

This set-up is conceptually closer because like SIMULLDA, it uses a structured interlingua. So the English acception river maps to an interlingual acception, to which no French acception is related. The French acception fleuve maps to a different interlingual acception, and the two interlingual acception are related at the level of the interlingua.

It is even hinted that the relation between the interlingual acceptions is similar to the relation between the interlingual meanings in the SIMULLDA set-up: *"[L]a sémantique d'acception* RIVER *correspond à l'union de sémantique des acceptions* RIVIÈRE *et* FLEUVE.*"*[14](Sérasset, 1994 [134]) This sounds very much like the analysis in the SIMULLDA set-up, for also there, we have that $\text{RIVER}' = \text{RIVIERE}' \cap \text{FLEUVE}'$.

But there are two important differences: the first is that relations between acceptions are only introduced when such a relation is needed to 'fill' a lexical gap. So overall, the set of interlingual acceptions is not structured, only those acceptions participating in a lexical gap are related to others. And more importantly, there is no notion comparable to that of a definitional attribute: although the interlingual acception FLEUVE is labelled as more specific than the one related to RIVER, there is no indication *why* it is more specific. Without a notion like that of definitional attributes, it is even hard to give a clear interpretation what it means for the acception RIVER to be the intersection of the acceptions RIVIÈRE and FLEUVE.

The other additional approach I would like to mention is the *ontology clustering* set-up proposed by Visser & Tamma (1999). Also this approach

[14][T]he meaning of the acception RIVER corresponds to the union of the meaning of the acceptions RIVIÈRE and FLEUVE.

has something like a structured interlingua: a shared ontology and attributes over the concepts in it. The similarity is most interesting in that Visser & Tamma describe a method for translating non-corresponding notions, consisting of seven steps. These steps conceptually resemble the lexical gap filling procedure: a translation is derived from a comparison between the 'interlingual meanings' that the two non-corresponding words express: *"the attributes of the concept in the source ontology are compared with the attributes of the hypernym [found in the shared ontology] to select the distinguishing features."* (Visser & Tamma, 1999 [12]). This approach shows that a solution similar to SIM*u*LLDA can be reached from a completely different point of view. However, it is difficult to really compare the clustered ontology approach to the SIM*u*LLDA set-up . At least from the point of generating translations, this approach is not worked out in sufficient detail. At least one point at which this appraoch differs from the SIM*u*LLDA set-up is that the structure of the shared ontology in the clustered ontology approach has to be generated by hand, instead of being the result of a structuring formalism.

From the analysis and comparison in this subsection, we can conclude that not only is it possible to get a multilingual alignment using the SIM*u*LLDA concept lattice for the case of river and fleuve, but the analysis in SIM*u*LLDA as given in figure 4.4 has definite advantages over competing analyses. This is not to say that the analysis in figure 4.4 is the only analysis that could be given in SIM*u*LLDA. This just concerns four of the many words in appendix A. Let us now turn to the rest of the words to get a more complete picture of the multilingual alignment in SIM*u*LLDA.

Other Bodies of Water

Given the fact that the discussion of the interlingual alignment of river, fleuve, rivière, and stream took over five pages, it is obvious that a full discussion of all the words in the appendix would take up too much space. Furthermore, a complete analysis would only be fruitful if it would give a complete and correct picture of the lexical field. But there is no guarantee that the analysis will be either correct or complete. That it is not necessarily correct is nicely illustrated by the previous subsection: the analysis in figure 4.4 is only a possible analysis of the relevant lexemes. It is even in conflict with the data presented in the monolingual dictionaries, and there is nothing assuring that it is the proper analysis.

That the analysis of the appendix does not have to be complete is because by the fact that it relates only to the appendix, the analysis would be based upon the words present in the appendix. This means that any lexical gap that is found is only a lexical gap within the boundaries of the words in the appendix. But there is no guarantee that the list of words for bodies of water in the various languages is complete: unlike a system like WordNet,

normal dictionaries do not yield complete sets of hyponyms. Therefore, the words in the appendix had to be acquired in some way. For this, the following method was used: first, all the hyponyms of water were sought in WordNet, the result of which is given in appendix A.1. These hyponyms were then looked up in bilingual dictionaries to find their respective translations. Since this would (probably) leave out exactly the lexical gaps, some other electronic tools, like EuroGlot, were consulted in order to find more related words in the various languages.

However, this process does not guarantee completeness. To give a concrete example: the word moat is not given as a hyponym in WordNet, probably because LDOCE defines it as the trench surrounding a castle, and not so much as the body of water this trench contains. The Dutch translation given by VDEN is gracht[15]. But gracht is not translationally synonymous to moat; it is a more general term. This suggests that moat is a lexical gap in Dutch. But actually there is a good translational synonym: slotgracht[16], which was not rendered by the method described. And it might very well be that similar cases still exist in the appendix. Furthermore, there is a certain arbitrariness in the listed entries, for instance in the fact that the French bas-fond and the Italian bassofondo are incorporated, but their English counterpart shallows or shoal is left out. These always concern words that indicate marginal cases of bodies of water.

Not all alignments are equally difficult. By just looking at the first few words from Italian, the differences in problems become very clear. The word bocca is defined in a way that is as close to identical as you can get to the way the English strait and the Dutch zeeëngte are defined: translated into English, all three simply read *narrow passage between two areas of land*[17]. But the Italian word baia, though presumably a good translational synonym of bay, is defined in a radically different way: as an *inlet of a sea or lake, large in the center and narrow at the mouth*.

To reduce the discussion of the appendix radically, I will leave out all discussion about terms that are presumably nice translational synonyms, and only focus on those terms that might be a lexical gap. There is of course no guarantee that no words were incorrectly assumed to be translationally synonymous. Also, I will only discuss words that have lexical gaps in English, and not the question of which of the English words might be a lexical gap in the other languages. And if various languages have translational synonymous words that are lexical gaps in English, they will be discussed only once. This has the disadvantage of being imprecise, and not giving an accurate count of the percentage of lexical gaps between the various

[15] Actually, it gives (wal)gracht, but walgracht is not given as a lexical entry by GVD, and was hence discarded.

[16] Although singel is also a candidate translation.

[17] Though it should be noted that in the Oxford-Ragazzini, strait is translated not as bocca, but as stretto or canale, and conversely the strait-meaning of bocca is completely absent.

languages, but it will give an idea about which additional formal concepts there will be, on top of those listed for the LDOCE words discussed in the previous subsection. These additional formal concepts, together with the ones stemming from LDOCE, should provide the entire multilingual concept lattice.

bief	part of a stream, between two waterfalls part of a canal, between two locks
bief	canal bringing water from a stream to a hydraulic installation
bisse	long irrigation canal taking water from the mountains to a cultivated terrain
calanque	straight, long creek, boarded by steep rocks (esp. in the Mediterranean)
gave	stream of water, torrent of the Pyrenees
seguia	irrigation canal in North Africa

Table 4.7: Lexical Gaps from French

All but one of the 8 possible lexical gaps from French are simple in terms of their SIMuLLDA analysis: they are simply (translational) hyponyms of English words, with one additional definitional attribute that the English lexicalisation does not take into account. For instance, English has (as far as I know) no word for the specific kind of stream that is dug to bring water to a hydraulic machine. French has a lexicalisation for this notion, with **to_hydraulic_machine** as definitional attribute.

Since these words all have additional attributes, there is also little doubt that they are in fact lexical gaps, at least with respect to the word in appendix A.2[18]. The words do not have lexical gaps in all other languages; Dutch does have words for the first two meanings of bief: rivierpand (or -vak), and kanaalpand, depending on whether it is a part of a river or a canal.

The only more difficult lexical gap is the word calanque. It is more difficult because of the following reason: the prototypical information *esp. in the Mediterranean* should not be taken as part of the definitional attributes (see section 3.3.2). And the rest of the definition of calanque is virtually identical to the LDOCE definition of fjord. But the English fjord and the French calanque are not translational synonyms; the English fjord is a translational synonym of the French word fjord. The reason for this is that the definitions for fjord in LDOCE and calanque in Petit Robert are both too imprecise. It is probably because of the presence of the word calanque that the Petit Robert gives a more precise definition of fjord: *ancient glaciated valley, invaded by sea*

[18]The English translational synonyms might exist, but absent because of the incompleteness described earlier.

water during deglaciation, characteristic of the sScandinavian and Scottish coasts. And a calanque is not an ancient glaciated valley, but an ancient riverbed. So the best solution is probably to adjust the English definition of fjord.

Bergsee	lake in the mountains
Gebirgsee	like in a mountain range
Gracht	navigable canal in Dutch cities
Rigole	deep stream, small canal for irrigation
Rinnsal	very small, slow moving water

Table 4.8: Lexical Gaps from German

Like the French calanque and seguia, the German word Gracht is a region-specific notion. Interestingly, it is a loanword from Dutch, but the Dutch word gracht is more general, and can also apply for instance to a moat. Dutch uses a compound for the more specific notion: stadsgracht (gracht of a city; city canal).

There is an additional issue with the words Bergsee and Gebirgsee. Firstly, they are defined in terms of different definitional attributes, which would make them not synonymous, but presumably they are. But secondly, they are compound nouns. The fact that the English translation (the compound mountain lake) is a multi-word unit because English happens to divide compounds more often by spaces, should not be a criterion in the question whether Bergsee and mountain lake are words. So probably, it would be incorrect to say that Bergsee does not have a lexical gap in Dutch since bergmeer also happens not to contain a space, whereas it is a lexical gap in English. But on the other hand, the answer to this question has little impact on the bilingual dictionary: the only effect is that the English translation might either be *mountain lake* or *lake in the mountains*.

singel(gracht)	canal around a city
slenk	puddle in the road
vaardiepte	depth in a waterway wrt its navigability

Table 4.9: Lexical Gaps from Dutch

Like the French fondrière, the Dutch word slenk is defined as a body of water; their English translation is rut, but a rut is only the track mark left by a wheel on a soft road.
The three Italian words affluente, immisario, and tributario are not all lexical gaps, though probably immisario is. Judging from the definitions, tributario should be a hyperonym of immisario and affluente, but strangely, it is explicitly indicated in the definition, that tributario is synonymous with affluente.

The Italian word rigagnolo has a sub-meaning which specifically ap-

affluente	torrent or river that jets its water into another bigger river
immisario	stream of water that ends in a lake or some other bassin
tributario	said of a river that flows its water into another river or lake, syn. affluente
rigagnolo	small stream of water, esp. those that run at the side of the street when it rains
torrente	short, steep, fast-flowing mountain stream, accessible for extreme high and low water levels

Table 4.10: Lexical Gaps from Italian

plies to water at the side of a street; this meaning in VDIN is indicated as a different sense, which has a lexical gap in Dutch; the Oxford-Ragazzini translates it as gutter, but a gutter according to some is the greto: the ditch the water is streaming in, and not the stream itself. So rigagnolo might be a lexical gap.

Russian is as a language (and as a culture) less closely related to English than Dutch and German. So one would expect more lexical gaps in Russian than in the Germanic languages. However, whereas the Germanic languages do show some lexical gaps, no such lexical gaps are to be found in appendix A.7. This might not be because there are no lexical gaps in Russian, but because lexical gaps are hard to find, and there are less electronic tools available for Russian to help find possible lexical gaps than there are for the other languages in the appendix. There are only two words that might be a lexical gap: *мель*, and *рукав*. But this is not on the basis of their definition, but on the basis of native speakers; a *мель* is a ford[19], but according to informants it is especially a place where sheep can cross a river; and *рукав* is an arm of a river, but unlike an affluent, it is most prominently a part of a river that splits off from rather than flows into the main river. So a *рукав* is a river, branching from another river and flowing into the sea. But since these meanings are not indicated in the Ozhekov definitions, I will leave them out nonetheless.

Within a multilingual setting, there is a possible troublesome case with dependent attributes. Take a word that has a lexical gap, such as the Dutch word singel (canal around a city[20]) in English. Now suppose there was a Dutch word singeltje, defined as *a small singel*, in other word, defined in terms of the definitional attribute **small <u>rest</u>(**SINGEL**)**. Since SINGEL has no

[19]It even has literally the same definition as the Dutch drecht: *crossable place in a river* – although a *мель* can also be a crossable place in another *водоём* (body of water).

[20]Also here, we follow the dictionary definitions; the definition stays the same even though the *Singel* in Amsterdam and Utrecht nowadays lay well within the city – they are now technically speaking *stadsgrachten* with the name *Singel*.

lexicalisation in English, this attribute is hard to lexicalise in English. This is resolvable, since one can use the lexicalisation *small for a canal around a city*, in which the generated definition for SINGEL is used. But the other problem is, that such definitional attributes will by definition never play a role in lexicalisations in English, so care has to be taken not to have this result in inappropriate lexical gaps.

4.2.3 Generating Bilingual Definitions

The purpose of the SIM*u*LLDA system, as explained in chapter 1, is to provide a multilingual lexical database, in which bilingual definitions for every pair of languages in the database can be generated on the fly. Therefore, the most important aspect of the empirical test is whether the lexicographic context which was informally described in this subsection, does in fact lead to correct translations for all 36 combinations of languages in the appendix. But it is not difficult at all, given that we have the results of the discussion in the previous subsection. Let me show this.

There are two kinds of cases: those words that do have a lexical gap, and those words that do not. The first case is easy to deal with. To take a relatively complicated example: the definitions of fjord in LDOCE and fjord in Petit Robert are very different. Still, they can be taken as translationally synonymous. That they are translational synonyms means that should share all their definitional attributes. What these attributes are is up to lexicographers to decide. They might be the attributes of ARM OF THE SEA, with the additional **narrow** **rest(**ARM OT SEA**)**, **between cliffs or steep slopes** and **found esp. in Norway**; they might be the attributes of VALLEY, with the additional **ancient**, **penetrated by sea water during deglaciation**, and **characteristic of Scandinavian coasts**; they might be a combination of the above, or simply other attributes. But whatever they are, they will be shared by fjord_{Eng} and fjord_{Fr}. That is to say, fjord_{Eng} and fjord_{Fr} will express the same interlingual concept, or mng(fjord_{Eng}) = mng(fjord_{Eng}). This means in turn that the lexical definition in French, generated for fjord_{Eng} will be fjord[21]. So for those words that have translational synonyms, bilingual definitions will not be a problem. As we have seen in the previous subsection, over 90% of the lexical definitions is not a lexical gap.

For the words that do have lexical gaps, the situation is slightly more complex. Take the French word bief as an example. One of the meanings it expresses is BIEF_2, which has no definition in English. This means that it is a lexical gap in English. However, BIEF_2 is a sub-meaning of CANAL, which does have a lexicalisation in English: canal. The definitional surplus is **qui conduit les eaux d'un cours d'eau vers une machine hydraulique**, which

[21] Which just means that they are translationally synonymous, which was the assumption, so this is hardly a surprising result.

is lexicalised in English by *bringing water from a stream to a hydraulic installation*. It follows that the definition that will be generated for bief will be *canal bringing water from a stream to a hydraulic installation*. This is exactly the translation of its definition in Petit Robert, as presented in table 4.7. This means that the definition in English for the foreign words that have lexical gaps in English will be precisely the definitions presented in tables 4.7 to 4.10.

4.3 Ships and Sails

As a final test, one more lexical field will be discussed here: words for sails on ships. In his thesis, van Campenhoudt (1994) discusses the different words for sails on a ship, as well as their multilingual treatment, at length. The names for sails on a ship have the advantage over, for instance, words for bodies of water that they name a fixed domain: sailing ships simply have sails that languages can name, so checking whether two words name the same sail is easy to check. The relevant discussion concerns the 30 square sails on a fully rigged 5-masted barque: 6 sails on each mast, and 5 masts on a ship. If we give the masts a number and the sails a name, the 30 sails will be labelled $1a \ldots 5f$ as depicted in figure 4.7.

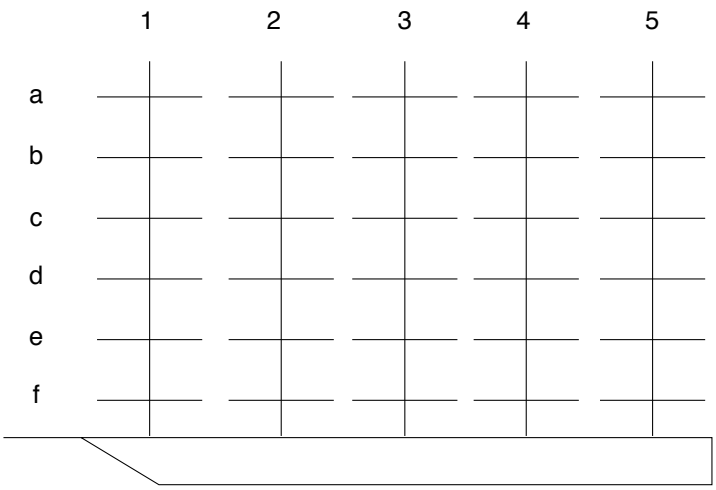


Figure 4.7: The Sails on a Ship (after van Campenhoudt, 1994)

All these different sails have a different name in French, for instance the name of 4c is *grand perroquet fixe arrière*. This naming is systematic, and consists of the composition of the name of the sail and the name of the

mast. The names of masts and sails in French, English, and German are given in table 4.11. There are some exceptions to this systematic naming: the 'mizen course sail' in English is called the crossjack, but most of the sails have their systematic names.

	French	German	English
1	petit	Vor	fore
2	grand avant	Groß	main
3	grand central	Mittel	middle
4	grand arrière	Kreuz	mizen
5	perroquet	Jigger	jigger
	perruche	Besahn	
a	cacatois	Royal	royal
b	perruche volant	Oberbramsegel	upper topgallant sail
c	perruche fixe	Unterbramsegel	lower topgallant sail
d	hunier volant	Obermarssegel	upper topsail
e	hunier fixe	Untermarssegel	lower topsail
f	basse voil	Untersegel	course sail

Table 4.11: Names for Sails and Masts in Figure 4.7

The words for these sails are not common words. To indicate just how uncommon they are: normally, the mast just behind the main mast (the highest one) is called the *mizen* mast (after the jigger mast can be the driver, the pusher, and the spanker mast), but only on a barque there is a *middle* mast in between, so these names only apply to 5-masted barques. And 5-masted barques are not very common, in fact only two fully rigged 5-mast barques have ever been built: the *Preussen*, built in 1902 by Reederei F. Laeisz (and run down in 1910, see figure 4.11). And in july 2000, the *Royal Clipper* was built by *Star Clippers*, inspired by the *Preussen*[22]. There were other 5-masted barques, such as the *France* (1890-1901), but they had less sails. So the complete terminology illustrated in figure 4.7 applies to just two ships.

The words for all these sails are predictable multi-word expressions. Therefore, few of them can be found in a normal dictionary like LDOCE or COD. Van Campenhoudt takes his data from a specialised dictionary: the trilingual *"de la Quille à la Pomme de Mât"* (from Keel to Truck) written by captain Heinrich Paasch (1885-1901). For a SIMuLLDA analysis, these data are not particularly well-suited: the dictionary written by Paasch is already constructed explicitly multilingually. Therefore, the comparison in this section will be based upon the data presented by van Campenhoudt rather than on existing dictionary data. This is in principle no problem: SIMuLLDA

[22] According to Star Clippers (pers. comm.), the middle mast on the *Royal Clipper* is indeed called the *middle mast*, although there are spelling differences: the 4th mast is called *mizzen mast*, and the 5th the *zigger mast*.

only aims at being true to dictionary data, not to build a database out of some particular existing dictionary.

If all languages have names for all 30 sails (as they do given their systematic names in table 4.11), there will be no lexical gaps. But the situation is somewhat more complex: not only do we have the 30 names for the 30 sails, but there are also hyperonyms. For instance, the French word grand perroquet can be any of the sails $\{2b, 2c, 3b, 3c, 4b, 4c\}$. German and English do not have such a hyperonym, since they have no concept grand mâtin the sense that French does. Also, in German, the aft mast has a different name depending on its rigging and the total number of masts: if it has gaffs, it is called the Besahn, while if it has yards, it is called either the Jigger (on a 4-masted ship) or the Kreuz (on a 3-masted ship). In French, the aft mast is called the perruche in all these cases. Both these differences lead to lexical gaps.

The way Dhydro deals with lexical gaps was already explained in section 1.2.2: for a word which only has translational hyponyms, the hyperonymic meaning is 'split up' into the more specific meanings of these translational hyponyms, thus linking the hyperonym to its translational hyponyms:

> [I]l convient d'adopter une démarche purement descriptive et donc de prévoir un mécanisme qui, en l'absence d'équivalent pour un même noeud du réseau sémantique, propose le choix d'un terme hyperonyme[23]. (van Campenhoudt, 2001)

How this works is best shown using an example. Because of the different naming of the aft mast, the topmost sail of the aft mast is always called *cacatois de perruche* in French, but in English it is called the *mizen-royal*, or the *jigger-royal*. So the French term cacatois de perruche has a lexical gap in English, and the English word mizen-royal has a lexical gap in French. van Campenhoudt (1994) represents this situation as in the upper part of figure 4.8: there are three notions, for each of which it holds that there either is no lexicalisation in French, or no word in English (Z stand for 'zero'). And this difference is due to the number of masts: the English terms apply either to a 3-masted ship with square sails (3MC) or a 4-masted square (4MC), while the French term relates to both (3MC&4MC).

In order to fill these lexical gaps, the English ambiguity is 'introduced into' French: the term *cacatois de perruche* is assigned to the notion related to both its translational hyponyms. This 'projecting down' is called *hyperonomase*, the result of which can be seen in the left bottom part of figure 4.8. After

[23] It is useful to take a purely descriptive stance, and to hence have a mechanism which, in the absence of an equivalent for a same node of the semantic network, proposes the choice of a hyperonymous term.

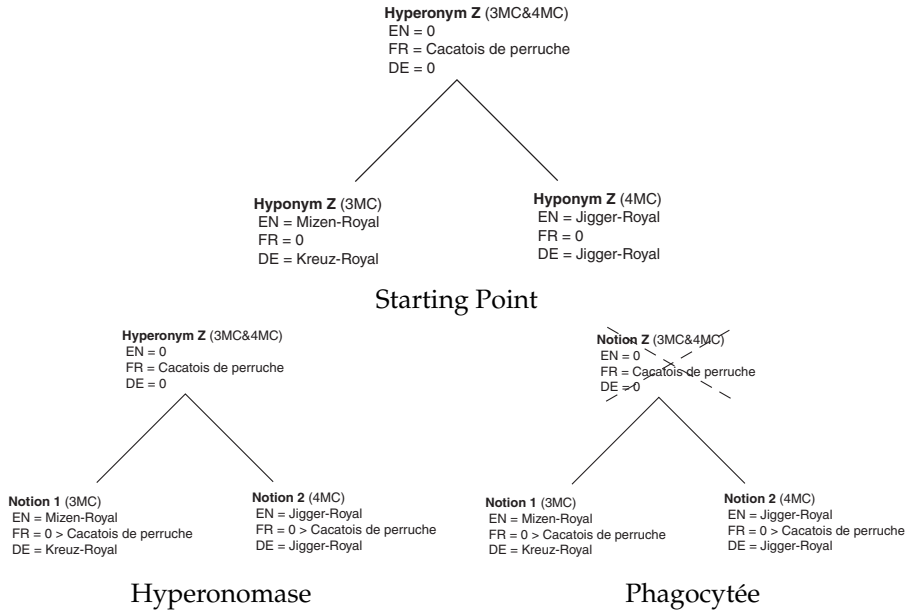


Figure 4.8: Hyperonomase and Phagocytée

hyperonomase, the lexical gaps for mizen-royal and jigger-royal have been filled. But hyperonomase leaves one term without a translation: the hyperonym Z. This hyperonym is now redundant, so it can simply be discarded. This discarding is called *phagocytée*. After hyperonomase and phagocytée, all lexical gaps have disappeared, as illustrated in the right bottom part of figure 4.8.

So in the Dhydro analysis, the term cacatois de perruche is not asigned its own meaning, but only the meaning of either the more specific mizen-royal, or jigger-royal. Notice that this is not specifically the choice of the Dhydro system, but a representation of the data as given by Paasch: on page 341 of his dictionary, Paasch (1901) simply gives meanings for cacatois de perruche as indicated by **notion 1** and **notion 2** in figure 4.8. The principles of hyperonomase and phagocytée simply provide a theoretical way of modelling these data.

In the SIM*u*LLDA set-up, both principles of hyperonomase and phagocytée are superfluous, since the starting point in figure 4.8 is already a good representation of the analysis that SIM*u*LLDA would give. This I mean in the following way: there are three interlingual meanings at play here: CAC_DE_PER, MIZEN_ROYAL, and JIGGER_ROYAL. These all share most of their definitional attibutes: they are all hyponyms of SAIL, located on the **aft_mast**, and attached to the **topmost_yard**. But on top of these, the meaning MIZEN_ROYAL has a definitional surplus over CAC_DE_PER: it also has

to be the sail of a **3-masted** square-sailed ship, and similarly JIGGER_ROYAL has to be on a **4-masted** square-sailed ship. So both are subconcepts of CAC_DE_PER, which itself is indifferent towards the number of masts[24].

In that analysis, the starting point in figure 4.8 would be a representation of a part of the concept lattice. And if we would add a French lexicalisation of the definitional attribute **4-masted** as `d'un 4 mâts carré`, the lexical gap filling procedure would nicely generate the French lexicalisation of jigger-royal: *cacatois de perruche d'un 4 mâts carré* (cacatois de perruche of a 4-masted square-sailed ship). Hyperonomase and phagocytée provide means to fill the lexical gaps in this starting point, but these are filled by other, more flexible means in SIM*u*LLDA.

To show just how nicely the analysis of the names of square sails on a 5-masted barque works in SIM*u*LLDA, the concept lattice of the entire set of names for topsails in French is illustrated in figure 4.9. The entire concept lattice for all square sails is larger, having slightly less than 3 times as many nodes.

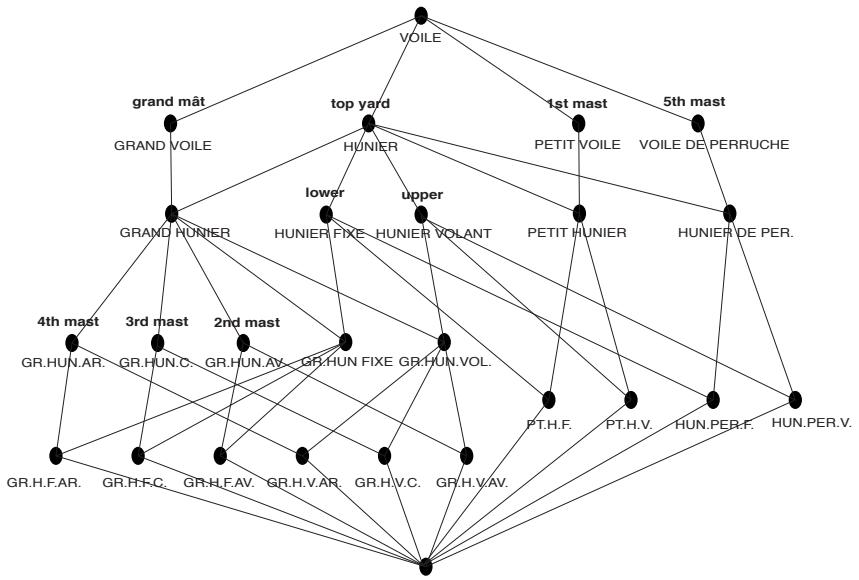


Figure 4.9: Concept Lattice for Topsails

There are a lot of nodes in the lattice in figure 4.9: 26 in total, while there are only 10 definitional attributes. But notice that for *every* node in the lattice (except $\perp$), there is a word in French. This illustrates nicely how strictly the names of the sails are combinations of the properties they have. Despite

[24]It could also be explicitly attributed **3/4-masted**, with specificity ordering on the attributes: **3-masted**, **4-masted** $\preccurlyeq$ **3/4-masted**.

these many names, the concept lattice correctly displays all hyperonymy relation between these terms.

The lattice in figure 4.9 also nicely illustrates the necessity to have a richer structure than a hierarchy: in a hierarchy, a choice would have to be made to first divide the sails according to their mast and than their yard, or the other way around. And in both these choices, a concept would have been missed: if the mast would be prompted, there would have been no concept HUNIER (topsail), whereas when the yard was prompted, there would have been no concept VOILE DE PERRUCHE (jigger sail). And because of this restriction, the names of sails are hard to capture in any hierarchy-based system, including for instance WordNet[25]

Notice that in the lattice in figure 4.9, there is a dependency between the attributes: the property **grand mât** is a weaker definitional attribute than either **4th mast**, **3rd mast**, or **2nd mast**. With the notion of partial order on attributes (see section 2.4.7), this can be modelled nicely.

If we try to map the English words onto this lattice, we find a number of lexical gaps, all of which are due to the fact that English has no notion of a grand mât, and hence no lexicalisations of the definitional attribute **grand mât**. There are 4 lexical gaps in English due to this: GRAND VOILE, GRAND HUNIER, GRAND HUNIER FIXE, and GRAND HUNIER VOLANT to be precise. Reversely, there are no lexical gaps in French for any of the English words for topsails. So the lattice in figure 4.9 is the complete interlingual lattice for these terms. The naming of the other square sails is isomorphic to that of the topsails, so the interlingual alignment of the square sails on a 5-masted barque can be completely correctly dealt with in SIM*u*LLDA.

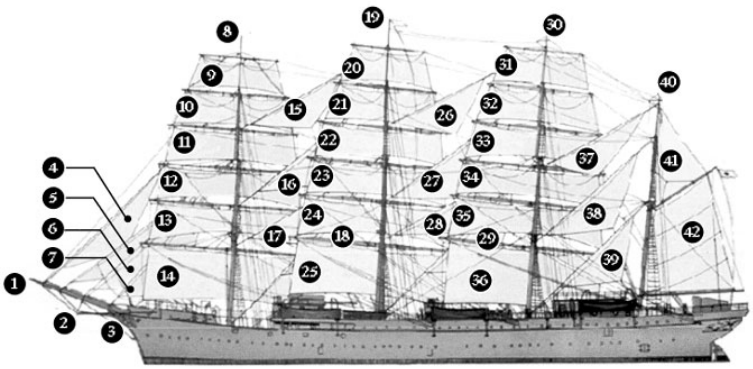


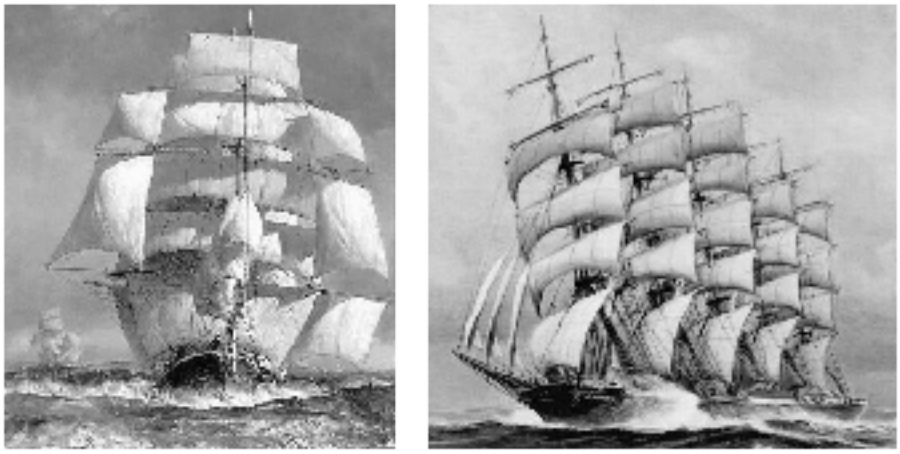
Figure 4.10: Illustration for Miami Herald, June 4, 2000

[25] Although WordNet could use the **part-of** relation to relate all the sails both to the mast they are part of, and the yard they are part of, but keep in mind that given the problem with **part-of** mentioned in 4.2.1, this is not immediately a solution.

Mizen-topmast-staysail	Diablotin	Kreuz-Stengestagsegel	3-masted square
Mizen-topmast-staysail	Diablotin	Besahn-Stengestagsegel	3-masted triangular
Jigger-topmast-staysail	Diablotin	Kreuz-Stengestagsegel	4-masted square
Jigger-topmast-staysail	Diablotin	Besahn-Stengestagsegel	4/5-masted triangular

Table 4.12: diablotin and its equivalents

The square sails are not the only sails on a ship; there are more sails, as is illustrated in figures 4.10 and 4.11. The additional sails come in three types: the *jibs*, which are attached to the bowsprit (4, 5, and 6 in figure 4.10), the *staysails*, which are in between two masts (15-17, 26-28, and 37-39 in figure 4.10), and the *studding sails* or *stunsails*, which are attached to booms attached to the end of the yards (on the very left and right of the Golden State, figure 4.11). The staysails and the studder sails can be attached to different yards and masts, so there is a *jigger topgallant staysail*, and a *fore royal studding sail*.



Golden State (1852, New York) Preussen (1901, Hamburg)

Figure 4.11: Great Historic Ships under Full Sail

These additional sails introduce additional definitional attributes, making the complete lattice for sails even bigger. And amongst the words for these additional sails, there are also lexical gaps. For instance, the French word diablotin is more general than the words in English or German, as is illustrated in table 4.12 (after van Campenhoudt (1994 [95])).

This example clearly shows a difficulty of the Dhydro approach: English and German both lexicalise a distinction that French does not make. And for both these distinctions, the process of hyperonomase duplicates the meanings of diablotin, resulting in 4 different meanings for this inherently monosemic word. If more languages were added, this can even lead

to an explosive number of meanings for diablotin. In the SIM*u*LLDA approach, the treatment of this example is straightforward: the word diablotin will be a translational hyperonym of all the four English and German words, simply kept apart by four definitional attributes.

So even though the names for sails on a ship are many and their relations are complicated, treating them in the SIM*u*LLDA approach is a straightforward process, resulting in precisely the number and kind of interlingual meanings that you would want, as nicely illustrated in figure 4.9.

4.4 Conclusion to Chapter 4

With the help of the empirical test in the present chapter, I hope to have shown that it is possible to apply SIM*u*LLDA to actual lexicographic data, that a SIM*u*LLDA analysis of the data presented in this chapter does yield the desired kind of results, and that the application of SIM*u*LLDA to these data has advantages over alternative systems such as EuroWordNet, Dhydro, or Conceptual Structures (Sowa).

The appliction of the SIM*u*LLDA set-up to actual dictionary data is not without difficulties: there are many complicated issues. The alignment across languages is a difficult issue, and some problems, such as the regular polysemy issue, even cannot be solved satisfactorily. But these difficulties are to be expected: getting English and French lexicographers to agree on the definition of fjord is no more difficult than getting different English lexicographers to agree on the correct definition of the English word fjord. Of course, the definition that is ultimately chosen is open for debate, but so is any definition in the monolingual English dictionary. So we can conclude that the application of SIM*u*LLDA is difficult, but less difficult than expected, and that the system is flexible and powerful enough to deal with the kind of problems that have arisen in this chapter.

Of course, the discussion in this chapter only tests a small part of dictionaries: the semantics definitions of entity nouns. That is not surprising, since the system was designed to deal primarily with these semantic definitions of entity nouns. However, in the next chapter we will look at some other components of dictionaries.

Chapter 5

Extending the System

The system as described and tested in the previous chapters is the core of the set-up of SIM*u*LLDA. Although it has not actually been implemented, the test in the preceding chapter shows that, as a multilingual lexical database, it is has clear advantages over alternative systems that have been implemented, such as EuroWordNet and Dhydro.

The purpose of this thesis, however, is not to provide a lexical database for semantic information on words, but to provide a system for generating (bilingual) dictionaries. Dictionaries do contain more than just semantic definitions: they contain collocations, labels, corpus examples, etc. Therefore, this chapter will discuss some proposals for how other dictionary components could be dealt with in SIM*u*LLDA: section 5.2 will discuss labels, examples, and collocations. Section 5.3 will deal with some other kinds of dictionary entries, mostly verbs, adjectives, and meaningful patterns. After that, some aspects of the application will be discussed, such as the question how to deal with huge amounts of data while still resulting in usable dictionaries will be addressed.

The first thing that will be discussed in this chapter is a proposal for an adjustment to something which has been discussed before: the treatment of inflections (and morphological derivations). The discussion of these additions is made possible (in part) by the fact that the different components of the SIM*u*LLDA set-up are kept apart carefully: strings, word-forms, lexemes, meanings, etc.

5.1 Derivations, Inflections, and Lexical Functions

Purely hierarchy (or network) based systems like (Euro)WordNet have a problem dealing with verb nominalisations. This was already discussed in section 1.2.3, but let me illustrate the problem again. Consider the definition in LDOCE for walker:

walker /ˈwɔːkə^{r}/ *n* **1** a person who walks, esp. for pleasure or exercise

The word walker is a a hyponym of person, and is so because of a regular morphological derivation involving the verb walk. This itself is not a problem. But the problem is, that walk is not the only verb that can be nominalised: many other verbs yield similar nominalisations. And since many of these nominalisations will be a hyponym of person, the meaning (or concept) PERSON will have a large amount of sub-concepts. These subconcepts cannot be distinguished by hierarchy alone; they are all direct hyponyms of person. So without further information, a hierarchical treatment of verb nominalisations is rather uninformative. SIM*u*LLDA solves this problem by having definitional attributes. These can model correctly that a walker is not just any person, but a PERSON **who walks**, hence distinguishing the word from all other kinds of person-words.

However, this does not mean that the analysis of walker in terms of PERSON and **who walks** is ideal; there are some major disadvantages. These disadvantages are related to the fact that, although the definitional attribute **who walks** contains the word *walks*, it is not actually related to the verb walk present elsewhere in the system: definitional attributes are atomic, unanalysed elements. In practice however, the relation between walker and walk is the basis of the definition of walker: the definition in LDOCE does not really 'define' the word walker, it merely indicates it as the nominalisation of the verb walk. For walker, as well as many other derivations (like the regular adverb to an adjective, the adjective of a noun, the -able form of a verb, etc.), the meaning of the derived word is defined by explaining the kind of derivation it is, and the word it is a derivation of.

This general problem expresses itself in several shortcomings of treating derivations unrelated to their stem. Here are four such problems:

1. It is not as efficient as it could be: storing the definitional attribute by linking it to another definition would be more efficient; not only in term of storage and computation of the data, but also in terms of compilation of the dictionary. Such derived definitions could be generated automatically by default, and need only be checked by lexicographers afterwards, instead of being produced manually from scratch.

2. It does not account for the productivity of these derivations. By modelling such derivations as regular morphological products, the deviant forms could be more clearly indicated. Also, it would provide the possibility to find the derivation by looking for its stem, in this case find the nominalisation starting from the verb. For foreign language learners, it would be useful to know that the -able form related to extend can be any of the forms extensile[1], extensible, extendible, and extendable.

[1] extensile is only correct in American English.

3. It does not account for the fact that the ambiguity of the verb is (in principle) inherited by its nominalisation; the fact that smoke has various senses means that smoker may relate to any of these senses — in *eel smoker* it will be the sense *to preserve and give a special taste to (meat, fish, etc.) by hanging it in smoke*, while in most other cases it will be the sense *to suck or breathe in smoke from (esp. burning tobacco, as in cigarettes, a pipe, etc.)*[2].

4. It does not account for the fact that hierarchical properties of the stem inherit onto its derivations: the fact that the noun pony is a hyponym of horse, implies that the adjective pony is a hyponym of the adjective horse (implying in turn that a pony ride is a kind of horse ride)[3].

In the light of these problems, let me describe a framework in which such derivational information is modelled structurally. Apresjan *et al.* have proposed a system, called the Meaning ⇔ Text Theory, which contains *lexical functions*. These lexical functions can be used to give the desired structural definition for such derivations: *"Not all the entry words are supplied with...definitions. Some words stand in such meaning relations to other (key) words as are regular and can be expressed through lexical function. In these cases a reference to the key word formulated in terms of lexical functions turns out to be sufficient."* (Apresjan *et al.*, 1969 [5]). In the next subsection, I will explain the system of lexical functions, and after that I will explore how lexical functions could be integrated into the SIMuLLDA set-up, in order to see whether this solves the problems just mentioned.

5.1.1 Meaning ⇔ Text Theory

The Meaning ⇔ Text Theory (henceforth MTT) is a theory that can deal structurally with derivations. But the original motivation for the framework was a different one: the system aims primarily at dealing properly with collocations. Collocations are 'arbitrary recurrent word combinations'. The words that make up the collocation are (hence) related to each other. The relation between them is not a conceptual association: although spider and web are clearly related words, they are not words that typically appear within a two-words distance of each other[4], and hence do not form a collocation together. Collocations have to (also) have a textual proximity. Prototypical examples of collocations are: *give an answer, raise an issue, raving mad, stark naked, flock of seagulls*, and *gaggle of geese*.

[2]The fact that it most commonly means as cigarette smoker should be indicated differently. Such frequency effects will be discussed in section 5.4.

[3]There are problematic cases: we usually do not say that veal is a kind of beef, even though calves are cows.

[4]Although spider's and web do form a collocation.

The relation between the words of a collocation is often expressible in term of a functional relation: give is the word to indicate 'doing' an answer (the act of answering), the noun gaggle is the word for a 'multitude' of geese, raving is a word for 'very' mad. The dependency of one word on the other led Žolkovskij & Mel'čuk (1965) to introduce the notion of a *lexical function*. Lexical Functions express this functional dependency between two words formally: the fact that the word indicating a 'multitude' for seagull is flock is modelled by defining a function *Mult* which, when applied to seagull, yields the word flock. The notation for this is: Mult(seagull) = flock, where *Mult* is the function, seagull the argument of the function, and flock the value. In MTT, lexical functions that define collocators in this way are called *collocational* lexical functions. indexlexical function!collocational

There is also a second class of lexical functions that has nothing to do with collocations, but that indicate non-collocational associations between words. These are called *paradigmatic* lexical functions. An example is the function A_0, which takes a noun and returns the adjective expressing something related to that noun. For instance, the adjective expressing *'related to/of the sun'* is solar: $A_0(\text{sun}) = \text{solar}$. It is this second type of lexical functions that is relevant for derivations. Collocations will be discussed in section 5.2.1. indexlexical function!paradigmatic

Lexical functions are defined as (partial) functions over lexical units. Their formal description in MTT is as follows:

> Lexical Functions serve to describe, in a systematic and precise way, all of the restricted lexical coocurrences of L and most of its syntactic and semantic derivatives. A Lexical Function **f** is a function that associates with a given lexical unit L, which is the *argument* or *keyword*, of **f**, a set L_i of (more or less) synonymous lexical units – the *value* of **f** – that express, contingent on L, a specific meaning associated with **f**: $\mathbf{f}(L) = \{L_i\}$. (Mel'čuk, 1995a [199])

Lexical functions do not yield (single) words, but sets of words (possibly consisting of just one word). This allows for the possibility to have various possible words for expressing the same thing, such as the alternative *Able*-forms of extend mentioned earlier. So Able(extend) = {extensile, extensible, extendible, extendable}.

Since both the argument and the value of lexical functions are (sets of) words (though possibly of a different word class) the argument of a function can again be taken as the keyword for another lexical function. This means that (paradigmatic) lexical functions can be stacked. To give an example: $A_0(\text{sun}) = \{\text{solar}\}$. But the word solar again can be the input of another lexical function, for instance $\text{Adv}_0(\text{solar}) = \{\text{solarly}\}$. And this can

also be done at once, by function composition: $\mathrm{Adv}_0(\mathrm{A}_0(\mathsf{sun})) = \{\mathsf{solarly}\}$[5].

It would be possible to combine this into a single lexical function, since this stack of lexical functions still relates two words (sun and solarly). However, doing so would not account for the fact that it is regularly and predictably possible to apply any appropriate lexical function to the argument of another.

Since lexical functions are viewed as (mathematical) functions, they also have an inverse; a function in which value and argument exchange their role. So $\mathrm{A}_0^{-1}(\mathsf{solar}) = \{\mathsf{sun}\}$[6]. So A_0^{-1} is the function yielding the noun related to an adverb.

The claim made by MTT is that there is a limited number of functional ways in which words can be related, and hence a limited number of lexical functions. Although the precise number of them is open to debate, Mel'čuk (1995b) assumes that there are 56 of them[7]. A number of these is illustrated in table 5.1.1, provided with some examples (the full list as given by the DECIDE project is given in section B.3 of the appendix).

Magn	The biggest degree	Magn(naked) = stark
Syn	Synonym	Syn(help) = aid
Conv	Converse term	Conv(contain) = be contained
Anti	Antonym	Anti(friend) = enemy
Gener	Genus term	Gener(blue) = colour
S_0	Nominalisation	S_0(move) = movement
A_0	Related adjective	A_0(sun) = solar
Adv_0	Related adverb	Adv_0(critical) = critically
V_0	Related verb	V_0(death) = die
S_1	First participant	S_1(employ) = employer
S_2	Second participant	S_2(employ) = employee
Mult	Aggregate/multitude	Mult(sheep) = flock
Sing	Single instance	Sing(news) = item
Cap	Head/chief	Cap(committee) = president
Equip	Staff/personnel	Equip(cloister) = monk
Centr	Center/culmination	Centr(crisis) = peak
Able	Potential	Able_1(burn) = combustible

Table 5.1: Some Common Lexical Functions (Apresjan *et al*, 1969)

[5]Formally, this is not entirely correct: $\mathrm{A}_0(\mathsf{sun})$ is a set, and hence not of the correct type to be the argument of Adv_0. So in stacking lexical functions, we will assume this to be a shorthand for: $\mathrm{Adv}_0(\alpha) = \{\mathsf{solarly}\}$, where $\alpha \in \mathrm{A}_0(\mathsf{sun})$.

[6]Here, the fact that A_0 yield a set is even worse; in fact A_0^{-1} is simply a new function, and not the mathematical inverse.

[7]Les functions lexicales standard simples sont au nombre de **56**.(Mel'čuk, 1995b [129]).

Examples of collocational lexical functions in table 5.1.1 are *Magn*, *Mult*, *Sing*, and *Centr*, which in table 5.1.1 yield the following collocations: *stark naked*, *flock of sheep*, *peak of a crisis*, and *news item*. Collocational lexical functions yield words that express some property of the keyword, and where the two words can be used together.

Examples of paradigmatic lexical functions are the semantic $Able_1$, *Syn*, and the syntactic V_0, and A_0. The paradigmatic lexical functions define syntactic or semantic relations between lexical items. The examples in table 5.1.1 define relations between burn and combustible, between move and movement, between help and aid, and between sun and solar. The relation between these words is not just any relation, but precisely the relation that is used to define the value of the lexical functions (combustible, movement, aid, and solar) in the dictionary[8]. This is illustrated in the following definitions from LDOCE:

movement /ˈmuːvmənt/ *n* **1** [C; U] (an example of) the act of moving or condition of being moved

combustible /kəmˈbʌstɪbəl/ *adj* **1** that can catch fire and burn easily

solar /ˈsəʊlər/ *adj* **1** of or from the sun

So lexical functions provide precisely the kind of structural definition for derivations that was argued for at the beginning of this subsection. In 5.1.3, I will show how lexical functions can be integrated into the SIM*u*LLDA set-up to arrive at the desired result. But before turning to that, let me first compare derivations treated in this way to the way inflections are 'stored' in lexemes in the SIM*u*LLDA set-up.

5.1.2 Lexemes as (Lexical) Functions

Inflected words, such as the tensed forms of verbs, and nominal declinations, are not listed as separate lemmata in SIM*u*LLDA. They are treated as part of a word-expression that is listed as a single lemma under its citation-form (see section 3.1). This mimics the dictionary practice, where plurals (for instance) are also a part of the definition of the headword. However, there are also cases where plurals are listed as separate headwords, as is in the definition for mice in LDOCE:

mouse /maʊs/ *n* **mice** /maɪs/ **1** (*often in com.*) a small furry animal with a long tail that lives in houses and in fields, related to but smaller than a rat

mice /maɪs/ *pl. of* MOUSE

There is a striking resemblance between this treatment of inflection and the treatment of derivations in dictionaries. Derivations can also be listed

[8]Fontenelle in his thesis even shows that lexical functions are sometimes coded even more explicitly in dictionaries: **-ery** *a collection of* (Fontenelle, 1997 [153]).

as separate lemmata, with little more than a reference to its root as definition (as exemplified in the previous section), or even incorporated in the lemma of its root. The only difference is that inflections are listed as grammatical information directly after the headword, whereas derivations can be listed as *run-ons* at the end of the definition. Examples of run-ons are grumpily, grumpiness, and renovation in the following dictionary entries from LDOCE:

> **grump·y** /ˈgrʌmpi/ *adj infml* bad-tempered and tending to complain: *She's very grumpy when her tooth aches.* **– -ily** *adv* **– -iness** *n* [U]
>
> **re·no·vate** /ˈrenəveɪt/ *v* [T] to put back into good condition by repairing, rebuilding, etc. *The old building is being renovated.* **– -vation** /ˌrenəˈveɪʃən/ *n* [C;U]

Technically, run-ons are considered separate lemmata, not different from normal, full entries, and only presented as run-ons (i.e. without an explicit definition) to save space. However, given the similarity between inflection and derivation, it is not surprising that various lexicographers have argued that we should extend the lemma itself to include derivations, and treat derivations on a par with inflections:

> [E]ntries tend to focus on similarities of form rather than on similarities of meaning. For example, strong is the normal lemma for stronger and strongest, and usually also the entry-word for strongly. But strength, strengthen, strengthening, etc. are often other entries, without any obvious links with the first, and the entries will be all the more separated if the dictionary is a big one. Theft is the lemma for thieves, but is not always linked to theft. Scholfield (1979) even argues that the entry for thief should also contain steal etc.
> (Béjoint, 1994 [193])

There are many words that are eligible for incorporation in the lemma in this fashion. All the words in table 5.2 are derived from the Dutch word kolonie (=colony). Although some are rather far-fetched, these could in principle all be listed as part of the entry for kolonie.

Notice that the meaning of all these derivations is, because of the structural way in which they are linked to the root, (almost) completely predictable from the meaning of the root, for instance, a kolonialisatricetje would be a *small female person who turns something into a colony* (or something along those lines).

So in a way, derivations and inflections can be treated on a par. However there is a problem: as in the case of the inclusion of steal in the entry for thief proposed by Scholfield, one could argue that lunar should be included in the definition of moon, being the value of A_0 (it is defined in LDOCE as *of, for, or to the moon*). But the adjective lunar had nothing to do with the word-form moon as such, only with one of its meanings: it has

kolonie (*n*)	kolonies	kolonietje	kolonietjes
koloniaal (*n*)	kolonialen	koloniaaltje	koloniaaltjes
koloniaal (*adj*)	koloniale	kolonialer	koloniaalst
kolonialist (*n*)	kolonialisten	kolonialistje	kolonialistjes
kolonialisme			
kolonialistisch (*adj*)	kolonialistische	kolonialistischer	kolonialistischst
koloniseren (*v*)	koloniseer	koloniseert	koloniserend
	koloniseerde	koloniseerden	gekoloniseerd
dekoloniseren (*v*)	dekoloniseer	dekoloniseert	dekoloniserend
	dekoloniseerde	dekoloniseerden	gedekoloniseerd
gekoloniseerd (*adj*)	gekoloniseerde	gekoloniseerder	gekoloniseerdst
gekoloniseerde (*n*)	gekoloniseerden		
kolonisatie	kolonisaties	dekolonisatie	dekolonisaties
kolonisator (*n*)	kolonisatoren	kolonisatortje	kolonisatortjes
kolonisatrice (*n*)	kolonisatrices	kolonisatricetje	kolonisatricetjes
kolonialiseren (*v*)	kolonialiseer	kolonialiseert	
	kolonialiseerde	kolonialiseerden	gekolonialiseerd
gekolonialiseerd (*adj*)	gekolonialiseerde	gekolonialiseerder	gekolonialiseerdst
gekolonialiseerde (*n*)	gekolonialiseerden		
kolonialisatie	kolonialisaties		
kolonialisateur (*n*)	kolonialisateurs	kolonialisateurtje	kolonialisateurtjes
kolonialisatrice (*n*)	kolonialisatrices	kolonialisatricetje	kolonialisatricetjes

Table 5.2: Extended Lexeme for *Kolonie* (colony) in Dutch

nothing to do with the poetic meaning of the word moon as indicating a month. So if such derivations are to be included in the word-expression, care should be taken not to connect them incorrectly.

5.1.3 Derivation and Inflection in SIM*u*LLDA

The words in table 5.2 are a mixture of derivations and inflections. It turns out to be hard to distinguish the derivations from the inflections in a clear way. As said before, the proposal is to treat derivations and inflections on a par in SIM*u*LLDA. This means on the one hand that derivations will have to be treated like inflections: they will be seen as part of the word-expression represented by the citation-form.

On the other hand it means that inflections will be treated as derivations in that they will also be labeled by lexical functions. All inflected forms in a lexeme should somehow be marked in order to establish which form it is. Lexical functions provide a good means of doing this. Appropriate *inflectional lexical functions* should be introduced for that purpose, as they are not a standard part of the repository of lexical functions. The set

of inflectional functions will have to be made dependent on the language. For instance for English, there should be a function *Plur* for the plural form of a noun (Plur(horse) = {horses}), and Past for the past tense of verbs (Past(go) = {went}). For French, there should be more inflectional functions for verbs, such as: Fut1p(arriver) = {arriverons} (to arrive). The argument of these inflectional functions will always be the citation-form of the lexeme[9].

The set of lexical functions used in SIM*u*LLDA will not contain the standard lexical functions for synonymy (*Syn*) and hyperonymy (*Gener*), or a lexical function for hyponymy ($Gener^{-1}$), since these will be modelled not at the level of the word-expression, but at the level of the interlingual meanings. So the set of lexical functions used in SIM*u*LLDA will both be larger and smaller than the standard set of lexical functions in MTT.

By the analysis of derivations as part of the lexeme, derivations will be directly linked to the citation-form of the lexeme. But as observed at the end of the previous subsection, derivations are not as tightly related to word-forms as inflections are. Whether or not a derivation belong to a word-form depends on the intended sense of the word-form: lunar is only an adjective belong to moon in its sense of the celestial body, and moons is only the plural of moon in its meaning of a month. So derivations seem to be related to word-senses rather than lexemes and citation-forms.

But this does not imply that it is incorrect to relate derivations to lexemes: as defined in the previous chapter, the same word-form can take part in (or even represent) various word-expressions or lexemes. So if lunar is taken as part of the lexeme *moon*, this implies automatically that there will be a distinction between two different word-expressions: moon-lunar for the celestial body, and moon-moons for the month.

With the use of lexical functions for derivations, the word law will be represented in SIM*u*LLDA as illustrated in figure 5.1, with inflections and derivations both marked by means of lexical functions (in the black boxes).

It is an open question derivations should exactly be incorporated in the word-expression in this fashion: although MTT does not have such a lexical function, it would even be perceivable that we define a function *Meat* to deal with the regular polysemy between an animal and its meat (and the other regular polysemy examples in table 3.6); so *Meat*(cow) = {beef} and *Meat*(horse) = {horse}. A European project called DECIDE[10] which used lexical functions to encode collocations, introduced a number of additional lexical functions, as illustrated in table 5.3.

[9] Although it could also be modelled more modularly, by having functions for the tenses and the persons separated: Fut(rire) = {rirai} (I will laugh), and 2Plur(rirai) = {ririez} (you will laugh).

[10] Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora

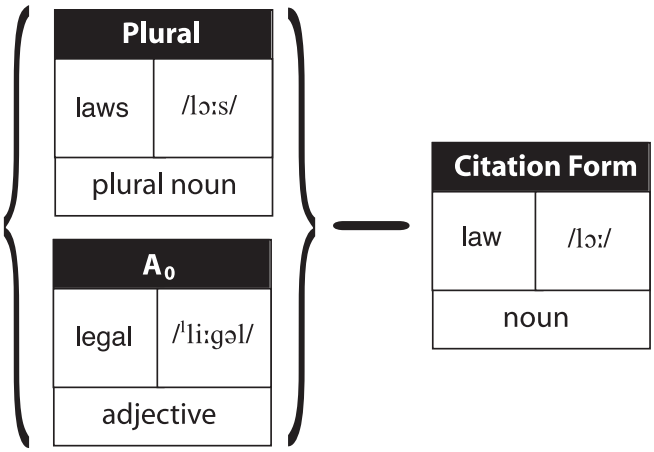


Figure 5.1: Word-Expressions in SIM*u*LLDA

Unit	Unit of	Unit(gravity)={g}
Part	Part of	Part(table)={leg}
Child	Child/young of	Child(horse)={foal}
Parent	Parent of	Parent(lamb)={sheep}
Male	Male of	Male(pig)={boar}
Female	Female of	Female(elephant)={cow}
Process	Process Noun	S_0Process(apply)={application}
Telic	Instrument/Verb	Telic(rubber)={erase}
Spec	Specific (hyponym)	Spec(flower)={rose, tulipe, . . . }

Table 5.3: Additional Lexical Functions of the DECIDE Project

These additional lexical functions are interesting, since with these lexical functions, it is in principle possible to define all the words in the concept lattice for horses in figure 2.5 in terms of lexical functions:

Female(horse) = {mare}
Male(horse) = {stallion}
Child(horse) = {foal}
Female(Child(horse)) = {filly}
Male(Child(horse)) = {colt}

The definition of filly in this fashion is an alternative way of characterising its meaning. There is a difference between this characterisation of filly and its treatment with definitional attributes: by treating it with definitional attributes, we say that filly is a word on its own, whose meaning happens to be related to the meanings of horse. While by treating it with the lexical

functions *Female* and *Child*, we say that filly is a derivation of horse, that is having its meanings from the way it is related to horse. And in the case of filly the first analysis is clearly the proper one.

But we have also seen an example where the second analysis is more appropriate (at least in part): in section 4.1, it was observed that Italian dictionaries do not have an entry for puledra (filly). And that the reason for that is that puledra is the regular female derivation of puledro. So puledra would be more properly dealt with as part of the lexeme *puledro*, with Female(puledro) = {puledra}.

Treating puledra in this way would result in a treatment in which Italian dictionaries will not have a lexical entry for puledra, since there is no lexeme with that citation-form (as should be). But at the same time, the word-form puledra would no longer be linked to the interlingual meaning FILLY, and hence not appear as the translation for filly in an English-Italian dictionary. To solve this, we need to reconnect the word-form puledra somehow to the interlingual meaning FOAL and the definitional attribute **female**.

There are two ways of doing this. The first is to treat puledra not only as the female derivation of puledro, but also as the citation-form of a lexeme of its own. This is possible, since as we have already seen, the same word-form can appear in more than one lexeme. It would also be improper to assume a strict separation of citation-forms and derivations. For one thing it is not always clear which of a pair of words should count as the stem, and which as its derivation: it is both possible to say that derive is the verb related to derivation, or that derivation is the nominalised verb of derive[11]. But this solution has two disadvantages: firstly, it would make the word-form as part of the lexeme puledro redundant. And secondly, it would nullify the advantage for the dictionaries with Italian as their source language.

An alternative way is to link the lexical function *Female* to the definitional attribute **female** and change the lexical gap filling procedure accordingly. The situation then becomes as follows: the English word-form filly expresses the interlingual meaning FILLY. And for FILLY, there is no lexicalisation in Italian (that is, there is no lexeme with puledra as its citation-form, which is related to the interlingual meaning FILLY). So we have to take the superconcept FOAL, and its Italian lexalisation puledro, together with the Italian lexalisation of the definitional attribute **female**. Now the default lexicalisation in Italian for **female** is `femmina`. But the definitional attribute **female** is also linked to the lexical function *Female*. This can be used to instruct the system to first check whether in the lexeme of the lexicalisation of the superconcept (*puledro*), there is a derivation with this lexical function. And there is: Female(puledro) = {puledra}. From this, we can conclude that puledra is the proper translational synonym of filly (if there would not have

[11] Dictionaries give answers to such questions: LDOCE does define derivation as *the process of deriving*, and not the other way around.

been such a derivation, the default lexicalisation would have been used).

So lexical functions can be used to model inflections and derivations structurally within the SIM*u*LLDA set-up. And it is even possible to relate these derivations indirectly, i.e. via their citation-form, to the interlingual concept lattice. The precise implications of this method would have to show in a larger empirical test.

5.2 Labels, Examples, Collocations

Although words and their meaning definitions constitute a crucial component of dictionaries, they are by no means the only information provided by dictionaries. Dictionaries also contain labels, collocations, and examples taken from corpora. In this section, I will review these different components, in order to establish whether or not they should find a place in the SIM*u*LLDA system; if so where, and if not, why not.

5.2.1 Collocations

As already mentioned in section 5.1.1, collocations form an important raison d'être for lexical functions. Collocations are best characterised in the following way: words can be put together to form phrases. Usually, the meaning of a phrase is a predictable combination of the meaning of the composing parts. Such phrases are called *free phrases*. However, some combinations like *red herring* are completely unpredictable and behave syntactically and semantically almost like a single word (*idiom*). In between free phrases and idiom are *collocations*.

There has been much discussion as to what collocations actually are. As said before, its generally accepted definition is the following: "*A collocation is an arbitrary and recurrent word combination.*" (Benson, 1990). So they are combinations of words that occur often, but not because of grammatical restrictions: the word on is often followed by the word the, because many NPs start with a definite determiner. But that makes it a rule-based, and hence non-arbitrary combination.

Collocations are belong to the class of phrasemes (phrasal expressions). There is a number of different phrasemes, with subtle individual differences. A classification of phrasemes by Mel'čuk can be found in figure 5.2. For present purposes, such a fine-grained classification of phrasemes is not necessary. Only the following observation is relevant: collocations allow more freedom than idioms, but they are less productive and predictable than free phrases, where free phrases and idioms are defined as follows:

Idiom A fixed combination of words, inside of which synonyms cannot be interchanged. So if kick is assumed synonymous with strike with the foot, it is still not possible to say *to strike the bucket with the foot*. Given

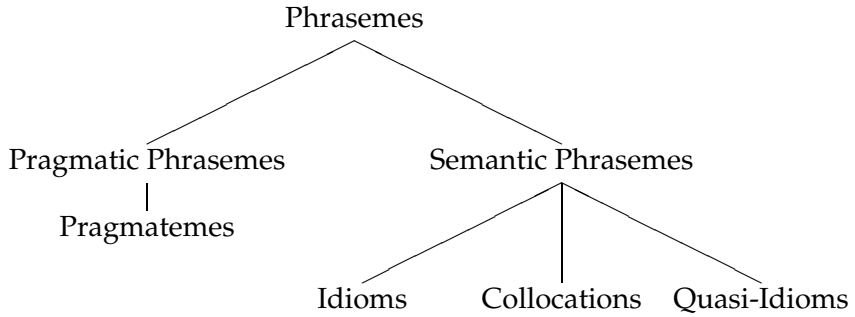


Figure 5.2: Classification of Phrasemes (Mel'čuk, 1995b [179])

the non-compositional structure of their meaning, idioms should be treated as multi-word lexemes.

Free Phrase A predictably and compositionally constructed word expression. Given their predictability, free phrases do not need to be incorporated in dictionaries.

Phrasemes are very numerous: "*In any language – i.e. in its lexicon – phrasemes outnumber words roughly ten to one.*" (Mel'čuk, 1998 [24]), and collocations form an important part of these phrasemes. Given their arbitrariness, collocations are also hard to learn: when learning English, you simply have to remember that you *make* a trip, but *take* a walk. Collocations are therefore not surprisingly the hardest and most erroneous part of second language acquisition, even in advanced stages. That is why collocations should not be absent from any good dictionary:

> Uppgifter om kollokationer är viktiga inslag både i enspråkiga och i aktiva tvåspråkiga ordböcker, eftersom användaren inte kan förväntas veta vilka ord som brukar uppträda tillsammans. I passiva tvåspråkiga ordböcker finns inte detta behov i samma utsträckning, eftersom kollokaten är givna redan i den text som skall tolkas. Däremot kan kollokationsuppgifter där ha en viktig uppgift när det gäller att leda användaren till rätt ekvivalent[12]. (Svénsen, 1987 [96])

The way collocations are present in for instance LDOCE is illustrated in table 5.4. There are two different ways of presenting collocations in this example: either without any indication of its meaning (raving lunatic), or

[12]Information about collocations is important in both monolingual and active bilingual dictionaries, since the user cannot be expected to know which words customarily occur together. In passive bilingual dictionaries, this need does not occur to the same extent, since the collocators are already given in the text that is to be understood, but on the other hand collocation can have an important function here as regards guiding the user to the correct equivalent.

with some small explanation (take a pulse). The reason why they can be given without explanation is the same as why they are less needed in passive dictionaries: they are often easy to interpret, though hard to produce correctly. So most or at least many collocations are only implicitly defined in the dictionary.

rav·ing /ˈreɪvɪŋ/ *adj, adv infml* **1** ralking or behaving wildly: *a raving lunatic* | *He's (stark) raving mad.*

pulse[1] /pʌls/ *n* **1** [*usu. sing.*] the regular beating of blood in the main blood tubes from the heart, esp. as felt at the wrist: *The doctor* **felt/took** *the woman's* **pulse**. (= counted the number of beats per minute) |*His pulse quickened/raced.* (= his heart beat very quickly) |*His pulse was strong/weak*

Table 5.4: LDOCE definition of raving and pulse

Because of their semantic transparency and their large numbers, it would be far from ideal (though not strictly speaking impossible) to treat all collocations as multi-word lexemes. Therefore, a more structured account of collocations should be given, and given the fact that SIM*u*LLDA already contains lexical functions, it is most logical to also use them for dealing with collocations.

Collocations in SIM*u*LLDA

Since lexical functions were designed to deal with collocations, treating them with collocations is in principle straightforward. The collocation *raving lunatic* can be dealt with by somehow having a relation between its two composing words, defined in terms of lexical functions. So one would expect the representation of *raving lunatic* to look like this: *Very*(lunatic) = (raving). But there is a problem with this analysis. The collocation *confirmed bachelor* does not relate to the word-form bachelor: it does not have anything to do with young seals, or people with a bachelors degree, both of which are word senses of the citation-form bachelor. If you use the turn of phrase *a confirmed bachelor* to refer to someone who refuses to get his masters degree, that would definitely count as creative word use, rather than the normal use of the collocation.

So if they do relate to words in a specific meaning, they could relate to interlingual meanings: *Very*(LUNATIC) = (RAVING). But this is even less true: in Dutch we say *poedelnaakt* or *spiernaakt*, which means *poodle naked* and *muscle naked* respectively. But in English we say *stark naked*. So since collocations are language-dependent, they can never be part of the interlingua. And they do not even relate to meanings within a language: in Dutch, there are several words for drunk: bezopen, zat, and lam. And all of these have different collocators: *starnakel bezopen*, *ladderzat*, and *stomdronken*.

So collocations relate to the 'things in the middle', i.e. the language-dependent word-senses. The problem is that word-senses are only im-

plicitly present in the SIM*u*LLDA set-up, by means of the relation between the citation-form and the interlingual meaning. They are the Saussurian concept of a word, referred to on page 67; pairs of form and meaning. Collocations will relate to pairs of citation-forms and interlingual meanings. For instance, the collocation *confirmed bachelor* will relate to the pair $\langle$bachelor, BACHELOR$_1\rangle$.

So collocations can be represented in SIM*u*LLDA by means of lexical functions on pair of citation-forms and interlingual meanings. And if collocations are hence present, they should also be provided with translational equivalents. This can be easily be integrated into the lexical definition generation process. To see how this can be done is best shown by using an example: following LDOCE, the English lexeme *bachelor* will be related (via its citation-form bachelor) to two interlingual meanings: BACHELOR$_1$ (MAN + **unmarried**), and BACHELOR$_2$ (PERSON + **with_bachelors_degree**). When generating a bilingual English-Dutch definition, SIM*u*LLDA will provide a translational synonym for bachelor in both these meanings, i.e. for both pairs $\langle$bachelor, BACHELOR$_1\rangle$ and $\langle$bachelor, BACHELOR$_2\rangle$. The second is a lexical gap in Dutch, and will by means of the lexical gap filling procedure be provided with the description *persoon met een bachelors diploma* (person with a bachelors degree); the first will give the citation-form vrijgezel.

Now on top of these word-senses, the first pair is also involved in a lexical function: Magn($\langle$bachelor, BACHELOR$_1\rangle$) = $\{\langle$confirmed, CONFIRMED$_1\rangle\}$. To find the translational equivalent of this collocation we have to find a similar collocation in Dutch. That is to say, we have to find a collocation represented by: Magn($\langle$x, BACHELOR$_1\rangle$) = $\{\alpha\}$, where x is a Dutch word-expression, and α is a pair of a Dutch word-expression and an interlingual meaning. And there is such a collocation, represented as follows: Magn($\langle$vrijgezel, BACHELOR$_1\rangle$) = $\{\langle$verstokt, CONFIRMED$_1\rangle\}$. So this procedure tells us that the English collocation *confirmed bachelor* has the Dutch translational equivalent *verstokt vrijgezel*. It can happen that there is no corresponding Dutch collocation. In that case, it would still be possible to indicate the meaning of the collocation, since the lexical function indicates what the relation between confirmed and bachelor actually is. In that case, either the lexical function itself could be given, or the lexical function could be assigned a general lexicalisation, like *very/strong* in English or *sterke/erge* in Dutch.

In this way, there is an elegant way of integrating collocations in the SIM*u*LLDA set-up, without having to define individual lexemes for them.

5.2.2 Corpus Examples and Illustrative Sentences

For many reasons, it is informative to see a word used in its actual context. Therefore, most dictionaries contain *illustrative examples*: sentences, mostly taken from some prestigious sources, in which the head-word of the dic-

tionary definition appears. For instance, the LDOCE definition of grumpy contains the sentence: *She's very grumpy when her tooth aches.* Such illustrative examples can serve a number of purposes:

1. *"They may be used in the dictionary to prove that a word or a particular meaning of a word exists in a language."* (Al-Kasimi, 1977 [89])

2. They can "*illustrate the grammatical (phonological, morphological, syntactic) behavior of the word defined in addition to their illustration of meaning.*" (Al-Kasimi, 1977 [90])

3. They can "*indicate – largely by the other words in them – something of the stylistic value of the entry.*" (Gleason, 1965 [429])

4. They can serve to illustrate the semantic range or distribution of the word.

Although illustrative examples are definitely illustrative, they are also problematic for many reasons. To start with, not all the purposes above are served by examples. With respect to the first purpose there is the following problem: simply naming a sentence in which the word is used only helps to prove its existence if its source is properly referred to. And there is no room in dictionaries to properly refer to the source. Also, a single occurrence can hardly be said to prove that the head-word is part of the language (see for instance section 3.2).

For the second and third purpose the limitation in size of dictionaries is similarly problematic: given the richness in structure of words, grammatical behaviour and stylistic value can only be illustrated very superficially by a single example. For both types of information, the dictionary entry contains items that are designed to better indicate that information: the grammatical label and the (register) label respectively.

The last purpose is not just badly served by illustrative examples, but using illustrative example for this purpose is even harmful for the quality of the dictionary. In fact, according to some, this is a reason to forbid or avoid the use of example sentences in some cases: "*To depend on quotations in this fashion ... is cheating on the part of the lexicographer who stops short of doing the descriptive work he ought to do*" (Al-Kasimi, 1977 [90]).

An option that is in many ways better than the use of illustrative sentences is using *corpus examples*. Corpus examples are not carefully selected sentences, but rather lists of places in a corpus where the word actually appears. They are not selected from, but rather linked to pieces of text, which makes it possible to see the word in its natural environment. Like illustrative sentences, they serve to show that the word is actually used, and illustrate its behaviour in terms of grammatical, semantic, and stylistic properties.

An example is given in table 5.5, where the COBUILD engine is used to select the first 40 occurrences of grumpy in the COBUILD Direct Corpus.

feel funny. Kind of pooped. Kind of	grumpy	." [p] He looked at her narrowly. ‘
was Judith Chalmers being a bit	grumpy	about a not quite inclusive enough
enjoys arguments; blames others;	grumpy	and morose; may simulate illness to
Letterman. [p] Appearing even more	grumpy	and foul-mouthed than usual, she
in his bed and would be there,	grumpy	and sarcastic, in the morning. She
St, our future landlord awaited,	grumpy	and red-haired, grudgingly
car was closing, and I felt as	grumpy	as ever. [p] Finally, I arrived.
this as a former baseball star	grumpy	at his demotion to manager/coach of
says [p] OR YOU might be left with a	grumpy	boyfriend, who, with no boys to
[/h] [b] Robin Young [/b] [p]	grumpy	British Father Christmases are to
was dry and hot, the hippos got very	grumpy	, but that has all changed. In fact,
she complained, made him tired and	grumpy	. But soon afterwards he married a
to lose than her rivals. [p] [h] Mr	grumpy	cheers Lizzie;Lizzie Power and
first come to the diner and the fat,	grumpy	cook. ‘One hamburger," Autumn said
complex woman, caught, briefly, on a	grumpy	day. In other circumstances I could
hooligans - but were met by a	grumpy	group of elderly choristers. [p]
s got all our money, he still gets	grumpy	if you don’t keep your tongue
word reaches me also that a dumpy,	grumpy	Irish veteran may be persuaded to
Taggart with his sick wife and	grumpy	Morse with his solitude, crosswords
This is far higher than anything the	grumpy	Mr Lamont has picked up showing
using when Oracle had been	grumpy	, ‘Oh, I’ll help if I can. I think
is definitely not easy for the	grumpy	old George Wilson (Walter Matthau).
years ago when nice Auberon Waugh’s	grumpy	old dad Beverley wrote an uncalled
of Wales giving an interview to	grumpy	old Panorama? Who could have
star was so miserable on the set of	grumpy	Old Men 2, her latest movie, that
get, a survey reveals. [p] And a	grumpy	one per cent complain they have to
street you would mistake him for a	grumpy	rabbi or some sort of lowly
backing onto an advert so that this	grumpy	sixty-three year-old novice
work and comes home late, tired and	grumpy	, so the last thing he wants is to
GLR’s Breakfast Show for being a	grumpy	sod, the 32-year-old Abbot learnt
remotely cost-effective, leading to	grumpy	staff and equally gloomy profits.
Leagues Under the Sea, in which the	grumpy	submariner is all sweetness and
lackeys or whoever? Do I hell! Mr	grumpy	, that’s me. I love it [p] Yes, but
a dark corner and labelled as too	grumpy	to play to anyone but the already
than last week. ‘She’s a little	grumpy	today," Barbara noted, adding that
talk to him," Snoot said, using the	grumpy	tone he’d fallen into lately. ‘What
CAN’T face going to church any more.	grumpy	vicar fixes us with his gimlet eye
by sipping champagne with TV’s	grumpy	Victor Meldrew. [p] On the day her
[p] On a more serious note,	grumpy	Victor Meldrew, alias actor Richard
at last. So do you," said Bear in a	grumpy	voice. You’re the one Arakny wrote

Table 5.5: COBUILD Concordancies for grumpy

Corpus examples are not without problems. For instance, to have a corpus example for every word in the dictionary, one needs an incredibly large corpus. Although the COBUILD corpus is large, with over 50 million words, there is no occurrence of a word like servomechanism, even though that word is common enough to appear in a medium-sized dictionary like LDOCE. But corpus examples serve all four purposes of illustrative examples better than illustrative examples themselves. And this is mainly due to the fact that they are not carefully selected. On the other hand, because corpus examples are not carefully selected, they are only useful in large numbers. And large numbers of examples are not fit for printed dictionaries. So corpus examples should only replace illustrative examples if the purpose of the lexical database is to generate electronic dictionaries.

5.2.3 Register Labels and Group Labels

As already mentioned in section 3.3.4, words that express the same interlingual meaning can still differ in important respects. In table 3.10, various words for policemen were given, that differ in that they are not all equally

formal. Such differences are usually indicated by *labels*. A few other examples of lexical entries with labels are listed in table 5.6.

cop[1] /kɒp||kɑːp/ *n infml* a policeman or policewoman
pissed /ˈpɪst/ *adj* [F] *taboo sl* **1** *BrE* drunk **2** *AmE* annoyed
op·pro·bri·ous /əˈprəʊbriəs/ *adj fml* (esp. of words) showing great disrespect

Table 5.6: Use of Labels in LDOCE

The role and treatment of labels is often a disregarded subject in lexicography. Let me here put forth the aspects which are most relevant for the SIM*u*LLDA system, and sketch how labels could be incorporated in the set-up of SIM*u*LLDA
footnoteIn the present subsection, I make use of Janssen *et al.* (to appear), a paper written on this subject..

Labels are mostly presented right at the beginning of the lexical entry, right after the grammatical information such as word-class and morphological information, often even having the same typesetting as the former. Yet there is a radical difference between wordclass indication and labels. Wordclass information is information about the word-form, a further specification of which word-form the definition actually defines. But labels primarily concern word-meanings; a word *in a specific meaning* may be taboo, slang, regional, or informal.

Within the set of labels, one may distinguish two different kinds of labels: *group labels* and *register labels*. A classification of labels is given in table 5.7. The function of these two kinds of labels is rather different (though there are intermediate forms, which makes it difficult to keep the two kinds of labels completely separate).

Class	*Subclass*	*Oxford labels*
Group labels Indicate words as belonging to group of speakers	Geographical	Afr. dial. north. Amer. *etc*
	Temporal	arch. mod. obs.
	Frequency	freq.
	Field	Aer. Alch. poet. techn. *etc*
Register labels Guide user in choosing between alternatives		colloquial, slang, jocular, derogative, vulgar, archaic, literary, euphemistic, figurative, pejorative suggested: very informal, informal, ∅, formal, very formal

Table 5.7: Classification and Examples of Labels (Janssen *et al*, 2001)

The primary function of group labels is to indicate that a word is not used uniformly throughout the entire community of a language, but that it is bound to a specific group or time. Group labels in a way indicate that the

word belongs to a dialect, be it a geographical dialect, the 'dialect' specific for a certain profession or social group, or a language that belongs 'in the past'. One could can say that group labels demarcate a language within a language (the more so because of the fuzzy boundaries between languages and dialects (see page 3.2.1). This is nicely illustrated by the fact that a regional word can be a lexical gap in the standard dialect. A good example is the Dutch regional word onk. It means odd in the sense that socks that are onk are lacking their other half. It is not a standard Dutch word, and is not even listed in the GVD, yet it is a perfectly good word in certain parts of the Netherlands.

Register labels on the other hand indicate words as 'marked', as having a certain connotation. They warn the user that the word should not be used in any context; a word that is marked *infml* should not be used in a business letter, for instance. This markedness is necessarily markedness relative to some other word, that has the same meaning, but a different connotation; the word *cop* is only informal because there is the word *officer* that has a neutral value. It would not make sense to provide a label *ifml* for the Dutch word huppelen (frolic) saying that it is an informal word, since there is no alternative word available.

Given the fact that register labels mark deviation from a neutral standard, there is always an implicit norm behind them; for instance, the labels could mark words that would not normally be heard on the BBC news. Not only is there an implicit norm, there is also an implicit graduality: words marked *taboo* are even less likely to be heard on the BBC news than those 'only' marked *infml*. Therefore, in Janssen *et al.* (to appear), we propose to make this situation explicit, by introducing scalar labels. On these scalar labels, words can be marked as deviating in either direction from an explicit norm. For instance, to take Dutch as an example, the words for 'making love' can be put on a scale, ranging from very informal (-2), via the neutral form (0, those words that could be heard on the Dutch NOS news), to very formal (+2). This scale is represented in figure 5.3. In the paper, the idea of a scale is given a wider application in the sense that figurative use can also be seen as having a -1 or -2 value w.r.t. a neutral 0 value for the nonfigurative use. In that sense, the scale is not tied up to five values.

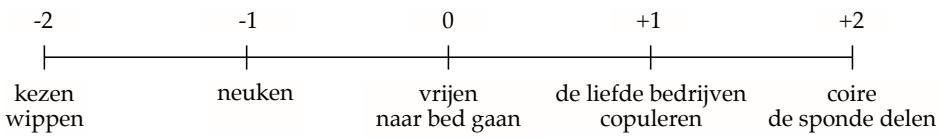


Figure 5.3: Scale of (In)formality for 'making love' in Dutch

Register labels give pragmatic information about the world; it is information quantifying over contexts in which the word is appropriate. Such kind of information is fundamentally different from semantic information about words.

Part of the philosophy behind SIM*u*LLDA is to closely consider the various components of dictionaries, and treat all different components at their appropriate level. This is why on page 94, it was argued that the colour of the word, i.e. its register label, should not be part of the meaning of the word. Also, meanings in SIM*u*LLDA are interlingual, as opposed to register labels by definition. Since labels quantify over linguistic contexts, they are by definition language dependent. That is why I propose to incorporate register labels at the level of the language-dependent word-meanings. That is, at the same level at which collocations are incorporated.

However, there is also good reason not to do this. One could equally well consider register labels to be interlingual, and treat them as definitional attributes (though be it definitional attributes of a very different kind). This would integrate labels more smoothly with the core system of SIM*u*LLDA, and it would also simplify an issue discussed before. On page 119, I discussed the dual approach dictionaries take towards the definition of synonyms. To incorporate this dual approach, the lexical definition generation procedure has to be adjusted to yield a different kind of definition for labeled and non-labeled word-forms. If we take labels as definitional attributes, this is not necessary: if briny is assigned the definitional attribute **lit or humor**, it will be a hyponym of sea, with *lit or humor* as the lexicalisation of the definitional surplus. In that case, the desired definition would be the default result of the lexical definition generator.

So in the SIM*u*LLDA set-up, it is very easy to treat register labels as (a special kind of) definitional attributes. Yet, it is methodologically more sound to treat them for what they are: markers of pragmatic aspects of the word-sense. Therefore, in spite of the appeal of doing things differently, I maintain that register labels should be dealt with at the level of language-dependent word-senses. So in SIM*u*LLDA, *lit or humor* will be an attribute of the pair ⟨briny, SEA⟩.

5.3 Abstract Nouns, Verbs, Adjectives

The discussion in this thesis has been focussed completely on a specific type of lexical entries: entity nouns, words denoting concrete objects in the external word. Entity nouns form a large part of dictionaries; they are by far the largest group of words. But they are of course not the only words, and a complete MLLD should also deal with other categories of words.

In principle, there is no problem in adding other kinds of words to the system. Abstract nouns, verbs, and adjectives can simply be analysed in

the same way as entity nouns:

love[1] /lʌv/ *n* **1** [U (**for**)]] (an example of) a strong feeling of fondness for another person, esp. between members of a family or close friends **2** fondness combined with sexual feeling

jog[1] /dʤɒg||dʤɑːg/ *v* **-gg- 3** [I] to run slowly and steadily, esp. for exercise

light[4] *adj* **1** of little weight; not heavy

So the word love can simply be analysed in terms of feeling, **strong**, **of fondness**, **for another person**, and **combined with sexual attraction**, and jog in terms of run, **slowly**, and **steadily**. Since interlingual meanings in SIM*u*LLDA are not denotational (see section 3.3.2), the fact that love, jog, and light do not denote objects is no restriction to treating them in terms of definitional attributes.

However, there are two points that should be considered. The first is, that a 'good' application of SIM*u*LLDA assumes that there is a certain layered hierarchy within words. So moussaka is a dish, which in turn is food, which in turn is a substance[13]. Without such a hierarchy, the SIM*u*LLDA set-up would loose much of its appeal. To see this, consider an extreme example, in which every word is defined as 'denoting an object of the related concept'. So for corkscrew we would have: *thing which is a corkscrew*. If this is taken as a lexicographic context, the corresponding concept lattice will be very flat: every formal concept will be a direct superconcept of $\perp$, and a direct subconcept of $\top$. That means that in this example, the interlingual lattice is very uninformative, and also the lexical gap filling procedure will not be very helpful: since there are no lexicalised superconcepts, every lexical gap will only be filled by the translation of the definitional attribute. So the SIM*u*LLDA set-up presupposes a certain amount of hierarchy amongst words.

With entity nouns, there is usually a nice hierarchy (although for natural language not as rich as biologists would have it). But verbs and adjectives usually display a less layered structure. Whether this lesser structure is still sufficient for a proper bilingual alignment as in the case of bodies of water discussed in the previous section is an open empirical question.

The second point is the following. Entity nouns have primarily a content: they serve to name (classes of) things. Do not have a lot of grammatical structure, which makes it easy for nouns to be linked across languages. Nouns can be count nouns or mass nouns, and where one language uses a count noun[14], another can use a mass noun, but for the largest part. But for the most part, comparing nouns of two languages concerns their meaning. For verbs this is different: verbs can be transitive, ergative, resultative, and

[13] After an example by Vossen & Copestake (1993).

[14] Verb nominalisations do of course have subcategorisation lists: you say *the gift of a book to John by Mary*. But as indicated in the previous section, these will not be treated using the interlingua at all, but by means of lexical functions.

perfective; they can have case markings on their internal objects, etc. So comparing verbs across languages involves a lot of grammar. And since the interlingual structure is based on the meaning of the words, this grammatical component of verbs can not easily be modelled in SIM*u*LLDA.

There are even word-classes with virtually no semantic context and a rich grammatical behaviour, such as prepositions and quantifiers. And these are therefore not very well fit for being modelled in the interlingual structure of SIM*u*LLDA. This does not have to be a large problem for two reasons: firstly, such word-classes are usually closed (there is a fixed and usually small number of them), and they are treated differently also in dictionaries, so there is no reason why they should be treated like entity nouns in SIM*u*LLDA.

So the SIM*u*LLDA set-up can be applied to more than just entity nouns; but depending on the behaviour of the word-class in question, the concept lattice itself might have less value for their definition.

There is an additional thing to consider. Languages do not always use the same word-class to express something. That might indicate that we would have to link nouns to other word-classes across languages, which apart from being undesirable in and by itself, would have the consequence of forcing all word-classes to be treated in the same fashion (or at least those word-classes that need to be cross-linguistically linked). An example where two languages use a different word-class can be found in the following two sentences, taken from Barnett (1994 [327]):

(5.1) I have a headache.

(5.2) Me duele la cabeza.

Although this is a much discussed problem in automatic translation, it is not so much a problem from a lexicographer's point of view. In an English-Spanish dictionary we simply find a nominal translation for the word headache: dolor de cabeza. So in bilingual dictionaries, there is no cross-category translation in such cases. The fact that the word headache is mostly used in the construction *having a headache*, with the above mentioned translation, is treated in a different way: by having a collocation *having a headache* under the lexical entry for headache, with the appropriate translation: doler la cabeza. But also in that case, two verbs are linked, so again there is no cross-category translation.

So for word-classes other than entity nouns in SIM*u*LLDA we can conclude the following: as far as their semantic content is concerned, there is no problem in treating them with a SIM*u*LLDA concept lattice in terms of definitional attributes. Whether this would lead to practical problems is an open empirical question. As far as their grammatical features are concerned, a different way of cross-linguistic linked should be used. So word-classes with a mostly grammatical character, such as prepositions, will have

to be treated by other means than the interlingual concept lattice. That this will lead to a different treatment of prepositions is not surprising: the lexical entries for prepositions in bilingual dictionaries are also very different from those for entity nouns.

5.4 Lexical Database, Size, and Word Frequency

As already observed in section 3.2.2, electronic dictionaries can contain much more information than their paper counterparts: a DVD could contain up to 13 million pages of typed text, which is far more than any dictionary every published, probably even more than all dictionaries taken together. This is in many respects a great advantage. However, there is a risk to massive expansion in size: the 'normal' user of a dictionary does not want to be flooded with unnecessary information. Compare the two definitions of the word chair in table 5.8, the first from the relatively small LDOCE, and the second from the much larger Oxford English Dictionary (second edition). The OED definitions are without the historical quotes (with them, the entry takes up almost 2 pages in the OED). For normal use, the massive definition in OED does not have added value, but simply makes it much more difficult to find the desired meaning of the word. The avarage dictionary user does not use LDOCE instead of OED not because it is cheaper or lighter, but because he does not need the lengthy definition in the OED.

So when a lot of data are present in the lexical database system, there is a need for a mechanism to perform a reduction of the outputted information, so that only a desired portion of the lexical information in the lexical database is rendered. This reduction involves three different things: the reduction of the number of lexical entries, the reduction of the number of senses presented for each lexical entry, and (possibly) the reduction of the definitions given for these word senses. Let me start with the first two aspects.

There is a very simple way to reduce the number of lexical entries, especially in a computational context. Take a large corpus, and for each word-form count how often it appears in the corpus. Then, for a dictionary with a desired number of lexical entries, determine the appropriate threshold, and yield only those lexical entries that exceed the threshold, so that exactly the desired number of lexical entries is given.

There are two problems with this method, both of which were mentioned before. The first is that counting word-forms in a corpus is in itself a difficult process: for homographs, it is hard to determine which of the word-forms is counted; it is hard (or at least currently not customary) to count all the inflected forms under the same number (since we are interested in the frequency of the lexeme, and not in the frequency of the

chair[1] /tʃeə^r / *n* **1** [C] a piece of furniture for one person to sit on, which usually has a back, a seat, four legs, and sometimes arms: *sitting on a chair at her desk—sitting in a comfortable chair watching TV* – see also ARMCHAIR, BATH CHAIR, DESKCHAIR, HIGH CHAIR, WHEELCHAIR **2** [C, *usu. sing.*] (the office, position, or official seat of) a CHAIRPERSON, esp. one in charge of a meeting: *Please address your remarks to the chair.—Who will be* **in the chair** *at tomorrow's meeting?—She's the chair of the house committe.* – see CHAIRPERSON (USAGE) **3** [C **(of)**] the position of PROFESSOR: *She holds a chair of chemistry in the university.* **4** [*the* + S] *infml* (esp. in the US) the punishment of death by means of an ELECTRIC CHAIR **5** [C] *old use for* SEDAN CHAIR

chair (/tʃɛə(r)/), *sb.*[1] Forms: 3 chaere, 4 cheiere, chaȝer, 4-5 chaier(e, chayer(e, 5 chaiare, chare, schayer, cheyer, cheare, chayr, 5-7 chayre, 6 cheyar, 6-7 chaire, 7- chair. [ME. *chaere, chaiere*, a. OF. *chaëre* (western and Anglo-Fr.), *chaierer* (= Pr. *cadera, cadeira*, Cat. *cadira*, OSp. *cadera*, Pg. *cadeira*): – L. *ca'tedra, cathedra* seat, a. Gr. καθεδρα, see CATHEDRA. *Ch-iè-re* was the regular OF. phonetic descendant of *ca-'ted-ra*; it was in Eng. also orig. of three syllables, afterward reduced to two *'cha-yer*, and finally (? under later F. influence) to one, *chair*. In the dialects it is still commonly of two, as Sc. *chayer* ('tʃejər). In mod.Fr. the phonetic variant *chaise* (see CHAISE *sb.*) has taken the popular senses, while *chaire* is restricted to the ecclestical or professorial *cathedra*.]

1. a. A seat for one person (always implying more or less of comfort and ease); now the common name for the movable four-legged seat with a rest for the back, which constitutes, in many forms of rudeness or elegance, an ordinary article of household furniture, and is also used in gardens or wherever it is usual to sit. **to take a chair**: to take a seat, be seated. **b.** With various substantives or adjs. indicating the nature, material, purpose, etc. as *bed-, bedroom, camp, cane, compass, folding, garden, hall, kitchen, leather, library, lobby, obstetrecal, office, rocking, swinging, Turkey, wheel-chair*, † *great-chair* (dial. *big-chair*, an arm-cahir. Also ARM-, BATH- (*sb.*[2]), CURULE-, EASY-, ELBOW-CHAIR. **c.** A glass-blower's sear furnished with long arms upon which he rolls the pontil; hence, the gang of men consisting of the glass-blower and his assistants. **d.** = *electric chair* (s.v. ELECTRIC *a.* 2). *U.S.*
2. *fig.* **a.** Seat. **b.** As an attribute of old age, when rest is the natural condition.
3. a. A seat of authority, state, or dignity; a throne, bench, judgement-seat, etc. **b.** *fig.* Place or situation of authority, etc. **c.** A chair occupied by a Welsh bard at an Eisteddfod, esp. one awarded as a trophy; also, a convention, now each of the four conventions, connected with the Welsh Eisteddfod.
4. a. The seat of a bishop in his church; hence *fig.* episcopal dignity or authority. *Obs* or *Arch.* † **b.** = SEE. *Obs.* † **5.** A pulpit. *Obs.*
6. a. The seat from which a professor or other authorized teacher delivers his lectures. **b.** Hence: office or position of a professor. **7.** A seat of judicial inquiry; a tribunal
8. The seat, and hence the office, of the chief magistrate of a corporate town; mayorship. *past, above* or *below the chair* (of aldermen of the City of London): having served or not served as Lord Mayor.
9. a. The seat occupied by the person presiding at a meeting, from whence he directs its business; hence, the office or dignity of chairman of a meeting, or of the Speaker of the House of Commons. **b.** Often put for the occupant of the chair, the chairman, as invested with its dignity (as *the throne* is for the sovereign), *e.g.* in the cry *Chair!* appealed to, or not duly regarded; *to address the chair, supposrt the chair*, etc. Now also used as an alternative for 'chairman' or 'chairwoman', esp. deliberately so as not to imply a particular sex. **c.** *pl.* The chairman and deputy chairman of the East India Company.
† **10.** An enclosed chair or covered vehicle for one person, carried on poles by two men; a sedan.
11. A light vehicle drawn by one horse; a chaise; also a particular kind of light chaise (see quot. 1795). Also *attrib.* **12.** *Railways.* † **a.** The support or carriage of a rail (cf. CARRIAGE 32 b). *Obs.* **b.** An iron or steel socket with a deep notch, into which the rail is fixed, and by which it is secured to the sleeper or cross-tie.
13. *Min.* (See quot.) **14.** Phrase. *to put in the chair.* (*slang*)
15. *Comb.*, as *chair-attendant, -back, -bearer, -bottoming, -caner, -caning, -cover, -factory, -hire, -leg, -maker, -mare, -mending, -room, -saddle, -slumber; chair-ridden, -shaking*, adjs.; **chair-back**, *(a)* the back of a chair; *b)* an anti-macassar; **chair-bard** [Welch *caddair fardd*], the successful competitor in the bardic competition held on 'chair day' of the Welsh National Eisteddfod; **chair-bed, -bedstead**, a kind of choair which can be unfolded into a bed; **chair binder** (see quot. 1921); † **chair-boll, chair bow**, a chair-back; **chair-borne** *a.*, *Mil.* ironically descriptive of troops whose duties are administrative; cf. AIR-BORNE *a.*, also *absol.*; **chair-car** orig. *U.S.*, a railway carriage furnished with chairs (two on each side of the aisle) instead of the usual seats; also , a parlour car (see PARLOUR 6); **chair-carver** (see quot.); **chair day**, the chief day of the Welsh National Eisteddfod (see quot.); **chair-days**, old age, when rest in a chair is the most natural condition; **chair-door** (see quot.); **chair frame maker, chair-framer** (see quots.); **chairlady** (orig. *U.S.*) = CHAIRWOMAN; *chair-lift* (see LIFT *sb.*[2]); **chair-marking** *slang* (see quots. and sense 14); so *chair-mark* v. and sb., *chair-marked* ppl. a.; **chair-matter** (see quot.); **chair-organ** (see quot.); **chair-post** *U.S.*, one of the main uprights of a chair; **chair-rail** (see quot.); **chair road**, a railway having the rails fastened by chairs to the sleepers; **chair rusher, seater** = *chair matter*; **chair-side**, *attrib.*, of or pertaining to dental work performed while the patient is seated in a dentist's chair; also in other uses (see quots.); **chair-table**, a table convertible into a chair or settle; **chair turner**, a wood turner who specializes in chair legs, rails, etc.; † **chair-volant**, sedan-chair; **chair-warmer** *slang*, orig. *Theatr.* (see quot. 1909); hence *gen.* a 'passenger' in any enterprise or situation. Also CHAIRMAN, *sb.*, etc.

Table 5.8: LDOCE and OED definitions of chair

citation-form); and the corpus has to be incredibly large to have sufficient occurrences of all the words.

The second problem is that frequency plays an important role in the decision which words to include in a dictionary of a given size, but it is not the actual criterion (as was discussed on page 77). There are various reasons why less frequent words might be selected over more frequent ones for a small dictionary. Therefore, it would be more appropriate not to use frequency as such, but to have a *saliency index*. Such an index is in fact a manual way to allow the lexicographer to decide at which level a certain lexical entry should be included: whether it should already be listed in a small dictionary, or appear in medium-sized dictionaries, or only be present in very large ones.

There is a problem with this solution. Consider the word rib, which is a common word, so it should get a hight saliency index. But the word rib does have various meanigs; it can also mean a kind af nitting pattern with a combination of plain and purl stiches producing a ribbed, somewhat elastic fabric (see page 3.5). And in that meaning (or sense), it is not common at all. So the problem is that it is in a way incorrect to assign saliency to lexemes.

The conclusion should be therefore that words are not the proper things to assign saliency indices to: the indices should be assigned to the word-senses (i.e. to pairs of citation-forms and interlingual meanings). The saliency of the lexeme is derived from the saliency of its word-senses. The most straightforward interpretation is this: all those word-senses that exceed the saliency threshold are yielded, so those lexemes for which at least one of the word-senes exceeds the saliency threshold will be listed. But it is also possible to use a more complex algorithm: when a lexeme is yielded because one of the word-senses exceeds the threshold, the threshold for the other word-senses of that word could be lowered. So the knitting-sense of rib might not be salient enough for a medium-sized dictionary, but since the word rib is present because of other senses of the word, it might be included anyway. Also, word-senses could combine their saliency to allow words with a number of almost-salient-enough word-senses to be listed. What the most appropriate strategy would be goes beyond this thesis.

In the description above, three different saliency indices were proper (for small, medium, and large dictionaries). But this number of saliency indices can be altered. The more indices there are, the more fine grained the selection of lexical entries can be. But on the other hand, the more indices there are, the harder it will be for lexicographer to consistently assign them. What the best number of indices would be is also an open question, and very much depends on the size of the lexical database.

Notice that with this method, every smaller dictionary is always contained in a larger one. This is not common practice in lexicography:

> Aucune nomenclature n'en contient une autre, elles ont toutes des

> entrées propres quelle que soit leur taille. Ce fait reflète des divergences dans l'appréciation de l'importance des mots assez courants, et non seulement des mots rares[15]. (Rey-Debove, 1971 [79-80])

To my knowledge, there is no principled reason why there should not be such a relation between smaller and larger dictionaries.

What is not accomplished easily in SIM*u*LLDA is the reduction of the definitions themselves: larger dictionaries often give more elaborate explanations about the meaning of the word. However, given the set-up of SIM*u*LLDA, in which lexical definitions are built out of definitional attributes and hyperonyms, it is hard to imagine how the same lexicographic context can be used to produce a more elaborate, or a more restricted definition depending on the size of the dictionary. Therefore, semantic definitions will not be reducible in the SIM*u*LLDA set-up.

5.5 Conclusion to Chapter 5

In this chapter, some possible adjustments and extensions to the system were discussed. A discussion of the treatment of some dictionary outside of the interlingua and Formal Concept Analysis, the structural treatment of derivations and inflections (with the use of lexical functions), the treatment of collocations (also with the use of lexical functions), corpus examples and illustrative sentences (with the use of concordancy software), and labels (with the use of scales and norms), and definition reduction (with the use of saliency labels).

With these additions, the SIM*u*LLDA set-up comes closer to a complete system. Still, there are various remaining issues that were not discussed in this thesis. Some of these issues will be touched upon in the next chapter, after the conclusion. But a full picture of all the problems and solutions will only emerge from an actual implementation of the entire system.

[15]No word-list is totally contained in another. Each, however small, has entries that no other has. This reflects the difference in assessment of the importanc of relatively common words, and not only of the rares ones. [translation by Béjoint (1971)]

Chapter 6

Conclusion and Afterthoughts

As outlined in the first chapter, the purpose of this thesis was a practical one: to construct a system (SIM*u*LLDA), 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 SIM*u*LLDA system, and describe whether it meets the requirements formulated above.

The basic lay-out of the SIM*u*LLDA 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 SIM*u*LLDA 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

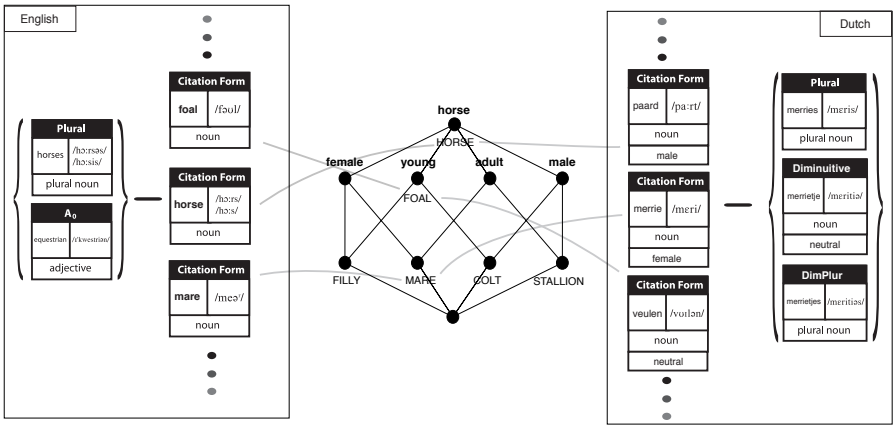


Figure 6.1: Set-Up of the SIM*u*LLDA System

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 SIM*u*LLDA 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.

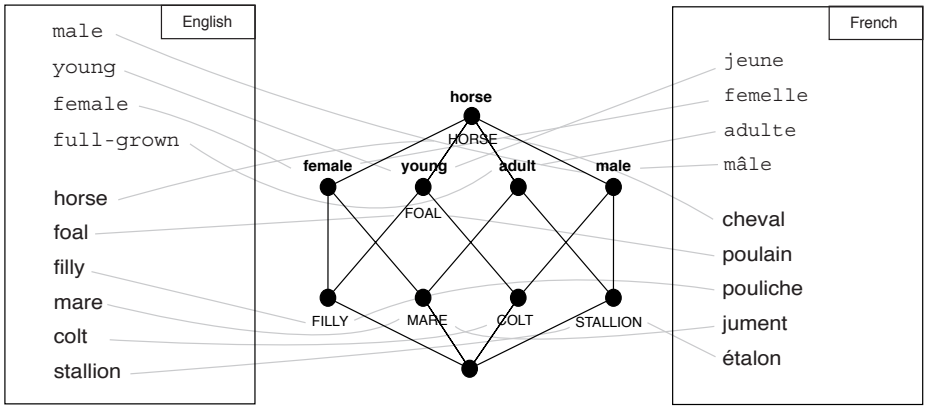


Figure 6.2: Lexical Gap in SIMuLLDA

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 SIM*u*LLDA 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 SIM*u*LLDA 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 SIM*u*LLDA 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 SIM*u*LLDA set-up and dictionary data, we can state that the SIM*u*LLDA 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 SIM*u*LLDA 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 SIM*u*LLDA. 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 SIM*u*LLDA 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 SIM*u*LLDA system puts some restrictions on the way in which lexical definitions can be given. This was avoided as much as possible, since SIM*u*LLDA 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 SIM*u*LLDA 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 SIM*u*LLDA 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 SIM*u*LLDA 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 SIM*u*LLDA 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.

Appendix A

Lexical Definitions for Water

A.1 'Water' in WordNet 1.6

results for *"Hyponyms (...is a kind of this), full"* search of noun *"water"*, Sense 2
body of water, water – (the part of the earth's surface covered with water)

- ⇒ drink – ((informal) any large deep body of water)
- ⇒ waterway – (a navigable body of water)
  - ⇒ mare clausum – ((closed sea) a navigable body of water under the jurisdiction of a single nation)
  - ⇒ mare liberum – ((free sea) a navigable body of water to which all nations have equal access)
  - ⇒ ditch – (any small natural waterway)
  - ⇒ rapid – (a part of a river where the current is very fast)
- ⇒ stream, watercourse – (a natural body of running water flowing on or under the earth)
  - ⇒ tidal river, tidewater river, tidal stream, tidewater stream – (a stream in which the effects of the tide extend far upstream)
  - ⇒ river – (a large natural stream of water (larger than a creek))
  - ⇒ brook, creek – (a natural stream of water smaller than a river (and often a tributary of a river))
  - ⇒ branch – (a stream or river connected to a larger one)
    - ⇒ feeder, tributary, affluent – (a branch that flows into the main stream)
    - ⇒ distributary – (a branch that flows away from the main stream)
  - ⇒ rivulet, rill, run, runnel, streamlet – (a small stream)
- ⇒ main, briny – (any very large body of (salt) water)
- ⇒ ocean – (a large body of water constituting a principal part of the hydrosphere)
- ⇒ sea – (a division of an ocean or a large body of salt water partially enclosed by land)
- ⇒ seven seas – (an informal expression for all of the oceans of the world)
- ⇒ high sea – (the open seas of the world outside the territorial waters of any nation)

- ⇒ territorial water – (the water over which a nation exercises sovereign jurisdiction)
- ⇒ deep, oceanic abyss – (an especially deep part of a sea or ocean)
- ⇒ mid-water – (the water that is well below the surface but also well above the bottom)
- ⇒ offing – (the part of the sea that can be seen from the shore)
- ⇒ lake – (a body of (usually fresh) water surrounded by land)
  - ⇒ reservoir, artificial lake – (a lake used to store water for community use)
  - ⇒ bayou – (a swampy arm or slow-moving outlet of a lake)
  - ⇒ loch, lough – (Scottish and Irish words for lake)
  - ⇒ lagoon, laguna, lagune – (a body of water cut off from a larger body by a reef of sand or coral)
  - ⇒ pond, pool – (a small lake)
    - ⇒ swimming hole – (a small body of water (usually in a creek) that is deep enough to use for swimming)
    - ⇒ fishpond – (a freshwater pond with fish)
    - ⇒ horsepond – (a pond for watering horses)
    - ⇒ mere – ((British) a small pond of standing water)
    - ⇒ millpond – (a pond formed by damming a stream to provide a head of water to turn a mill wheel)
    - ⇒ water hole – (a natural hole or hollow containing water)
  - ⇒ tarn – (a mountain lake (especially one formed by glaciers))
  - ⇒ oxbow lake – (a crescent-shaped lake (often temporary) that is formed when a meander of a river is cut off from the main channel)
- ⇒ shoal, shallow – (a stretch of shallow water)
- ⇒ gulf – (an arm of a sea or ocean partly enclosed by land; larger than a bay)
- ⇒ ford, crossing – (a shallow area in a stream that can be forded)
- ⇒ estuary – (the wide part of a river where it nears the sea; fresh and salt water mix)
  - ⇒ firth – (a narrow estuary (especially in Scotland))
- ⇒ waterfall, falls, cascade, cataract – (a steep descent of the water of a river)
- ⇒ cove, inlet, recess – (a small arm off of a larger body of water (often between rocky headlands))
  - ⇒ loch, lough – (a long narrow inlet of the sea in Scotland (especially when it is nearly landlocked) and in Ireland)
  - ⇒ fjord, fiord – (a long narrow inlet of the sea between steep cliffs; common in Norway)
- ⇒ bay – (an indentation of a shoreline larger than a cove but smaller than an gulf)
  - ⇒ bight – (a bay formed by a bend (a bight) in the shoreline)
- ⇒ sound – (a large ocean inlet or deep bay)
- ⇒ channel, sound – (a relatively narrow body of water linking two larger bodies)
  - ⇒ watercourse – (natural or artificial channel through which water flows)
  - ⇒ tideway – (a channel in which a tidal current runs)

- ⇒ strait – (a narrow channel of the sea joining two larger bodies of water)
  - ⇒ narrow – (a narrow strait connecting two bodies of water)
- ⇒ canal – ((astronomy) an indistinct surface feature of Mars once thought to be a system of channels; they are now believed to be an optical illusion)
- ⇒ rill – (a small channel (as one formed by soil erosion))

⇒ pool, puddle – (a small body of standing water (rainwater) or other liquid)

- ⇒ mud puddle – (a puddle of mud)
  - ⇒ wallow – (a puddle where animals go to wallow)

A.2 English Definitions

Bron:	**Longman Dictionary of Contemporary English**
bay[1]	a wide opening along a coast; part of the sea or of a large lake enclosed in a curve of the land
bayou	a body of water with a slow current and many water plants
bight	a curve in a coast larger than, or curving less than, a bay[1]
branch	**2** a separate and usu. less important part of something larger
briny	*lit or humor* the sea
brook[1]	a small stream
canal	an artificial waterway dug in the ground **a** to allow ships or boats to travel through **b** to bring water to or remove water from an area
cascade[1]	**1** a steep high usu. small waterfall, esp. one part of a bigger waterfall
cataract	**1** a large waterfall
channel[1]	**1** a narrow sea passage connecting two areas **2** the deepest part of a river, harbour or sea passage
cove[1]	a small sheltered opening in the coastline; small bay[1]
creek	**1** *BrE* a long narrow body of water reaching from the sea, a lake, etc. into the land **2** *AmE* a small narrow stream
crossing	**1** a place at which a road, river, border etc., can be crossed
deep[3]	*poet* the sea
ditch[1]	a V- or U-shaped passage cut into the ground, esp. for water to flow through
estuary	the wide lower part or mouth of a river, into which the sea enters at high tide
falls	a place where a river makes a sudden deep drop; waterfall
feeder	**2** a branch road, airline, railway line, etc. that connects with a main one

firth	a narrow arm of the sea, or place where a river flows out
fjord	a narrow arm of the sea between cliffs or steep slopes, esp. in Norway
ford[1]	a place in a river where the water in not very deep, and where it can be crossed on foot, in a car, etc. without using a bridge
gulf	**1** a large deep stretch of sea partly enclosed by land
high seas	the oceans of the world which do not belong to any particular country
inlet	**1** a narrow stretch of water reaching from a sea, lake, etc. into the land or between islands
lagoon	a lake of sea water partly or completely separated from the sea by banks of sand, rock, coral, etc.
lake[1]	**1** a large area of water, esp. non-salty water, surrounded by land
loch	**1** a lake **2** a part of the sea partly enclosed by land
lough	a lake or a part of the sea almost surrounded by land
mere[2]	a lake
millpond	**1** an area of water used for driving the wheel of a watermill
moat	a long deep hole, usually filled with water, dug **a** for defense around a castle, fort, etc.,in former times **b** round an area for animals in a modern zoo, to stop them from escaping
ocean	**1** the great mass of salt water that covers most of the Earth's surface **2** any of the great seas into which this mass is divided
pond	an area of still water smaller than a lake, esp. one that has been artificially made
pool[1]	**1** a small area of still water in a hollow place, usu. naturally formed **4** a deeper part of the river where the water is almost still
puddle[1]	a small amount of water, esp. tain lying in a hollow place in the ground
rapids	a part of a river where the water moves very fast over rocks
reservoir	**1** a place where liquid is stored, esp. an artificial lake to provide water for an area
rill	*poet* a small stream
river	a wide natural stream of water slowing between banks into a lake, into another wider stream, or into the sea
river basin	an area from which all the water flows into the same river
rivulet	*lit* a very small stream
runnel	*lit* a small stream

sea	**1** the great body of salty water that covers much of the Earth's surface; ocean **2** a large body of salty water smaller than an ocean, either **a** part of the ocean **b** a body of water (mostly) enclosed by land
shoal	an underwater bank of sand not far blow the surface of the water, making it dangerous to boats
sound[5]	**1** a fairly broad stretch of sea water mostly surrounded by coast **2** a water passage connecting two larger bodies of water and wider than a strait
strait[1]	a narrow passage of water between two areas of land, usu. connecting two seas
stream[1]	**1** a natural flow of water moving across country between banks, narrower than a river
tarn	a small mountain lake or pool, esp. in the north of England
territorial waters	the sea near a country's coast, over which that country has legal control and in which foreigners are not allowed to catch fish
tideway	**1** a narrow stretch of water through which the tide flows
torrent	a violently rushing stream, esp. of water
tributary[1]	a stream or river that flows into a larger stream or river
wallow[2]	**2** a place where animals come to wallow
wadi, wady	a usu. dry river bed in a desert, esp. in North-Africa
water[1]	**2** a mass or area of water, such as a lake, ocean, or river
water hole	a small area of water in dry country, where wild animals go to drink
watercourse	**1** a natural or artificial passage through which water flows **2** a stream of water, such as a river or underground stream
waterfall	water of a stream, river, etc. falling straight down over rocks, sometimes from a great height
waterway	a stretch of water, e.g. part of a river, which ships or boats can move on

A.3 Dutch Definitions

Source:	**van Dale Groot Woordenboek der Nederlandse Taal**
[4]**baai**	1 ronde inham van de zee in het land; kleine golf; kleine zee-boezem, syn. zeearm
beek	1 smal stromend water dat overal doorwaadbaar is
bergmeer	hoog tussen de bergen gelegen meer
binnenwater	1 niet in zee uitmondende stroom 2 polderwater

[1]**bocht**	3 buiging van een kust; – (vand.) golf, baai, inham
delta	2 land ingeloten door de armen waarin zich een rivier bij zijn uitmonding verdeelt
fjord	smalle, diep in het land dringende, zich vaak vertakkende inham met steile wanden in een bergachtige zeekust
drecht	2 doorwaadbare plaats in een rivier
[2]**golf**	wijde baai, ruime zeeboezem
gracht	2 met water gevuld kanaal, m.n. om of door een stad of rond een vesting 4 (gew) gedeelte van een hoofdgracht, tussen twee dwarsstraten of bruggen in
inham	1 in het land inspringende gedeelte van een zee, meer, rievier of ander (groot) water, kleine bocht of baai
haf	(aardr.) strandmeer, achter een landtong of enige kusteilanden gelegen inham van de zee, m.n. aan de kust van de Oostzee
kanaal	3 smalle natuurlijke verbinding tussen grote watervlakten (meren, zeeën) 4 kunstmatig gegraven waterweg (voor verkeer, afwatering of bevloeiing)
kanaalpand	afdeling, vak in een kanaal tussen twee sluizen
kanaalvak	gedeelte van een kanaal, syn. *kanaalpand*
kratermeer	met water gevulde krater
kreek	1 klein, smal, veelal stilstaand, niet gegraven water, dikwijls een inham van de zee, ook wel een overblijfsel van een overstroming of van de vroegere loop van een rivier; – smal vaarwater tussen ondiepten of eilanden 2 (gew) kleine rivier
lagune	door een lange, smalle landtong van de zee gescheiden klein strandmeer
[2]**maar**	**1** (gew.) gracht, afvoerkanaal **2** mare[3]
[3]**mare**	ketelvormige inzakking in niet-vulkanisch gesteente, gewoonlijk een meer
meer	binnenwater van enige omvang, m.n. een met water gevuld bekken
meertje	1 klein meer
modderpoel	waterplas met veel modder
oceaan	1 de grote wereldzee, de uitgestrekte en samenhangende water-massa die het land van de aardbol omspoelt
ondiepte	2 ondiepe plaats in een vaarwater
[1]**plas**	1 kuil met water dat na de regen niet in de bodem is getrokken 3 stilstaand water 4 grote watervlakte
[1]**poel**	1 klein ondiep, stilstaand water, syn. plas
ringgracht	gracht die een versterkte plaats geheel en al omringt

rivier	1 waterloop die door vereniging van beken of andere waterlopen op natuurlijke wijze ontstaat
rivierarm	tak van een rivier die zijn water uit de hoofdrivier ontvangt
riviermond	plaats waar een rivier in zee of in een andere rivier uitstroomt
riviertak	rivierarm
riviervak	groter of kleiner deel van een rivier tussen twee krommingen in
[1]**singel**	2 de gracht zelf om een stad
singelgracht	gracht om een stad
slenk	4 plas, gat in de weg
[1]**sloot**	1 gegraven water, smaller dan een gracht en breder dan een greppel, als afscheiding of om overtollig water af te voeren
slotgracht	gracht om een slot of versterkt kasteel
stadsgracht	**1** gracht om een stad **2** elk van de grachten in een stad
straat	zeeëngte
strandmeer	aan een kustvlakte gelegen meer
stroom	2 zich voortbewegende massa van een vloeistof, m.n. zich voortbewegende watermassa 6 door oevers of banken begrensd water dat langs een natuurlijke weg afvloeit; grote rivier
stuwbekken	bekken, bassin dat door een stuwdam wordt afgesloten, syn. stuwmeer
stuwmeer	1 stuwmeer
vaardiepte	diepte van een waterloop m.b.t. zijn bevaarbaarheid
vaargeul	1 geul van voldoende diepte als vaarwater tussen twee zandbanken of ondiepten door
vaarsloot	bevaarbare sloot, sloot die gebruikt wordt voor het verkeer te water
vaart	gegraven waterweg, syn. *kanaal*
vaarwater	1 zee die of gedeelte van een zee dat of zover het bevaren wordt 2 bevaarbare geul tussen twee gevaarlijke plaatsen door, syn. zeegat 4 binnenlandse waterweg die hoofdzakelijk dient voor de scheepvaart 5 bevaarbare geul in een binnenlandse waterweg
veenplas	door uitvening ontstane plas
ven	1 naam voor kleine meertjes, ook wel droge kommen 2 doo uitvening ontstane plas
[1]**vijver**	1 klein, natuurlijk of (meestal) gegraven, veelal omsloten waterbekken, m.n. in tuinen 2 (gew) open plas, syn. veenplas
wad	1 doorwaadbare plaats

water	3 genoemde vloeistof zoals zij voorkomt in haar natuurlijke of aangelegde bedding; alg. naam voor meren, rivieren, sloten, kanalen, enz.
waterbassin	waterbekken
waterbekken	natuurlijk met water gevulde kom, syn. bassin, meer
waterloop	2 stroom, wetering; beek
waterplas	natuurlijk, stilstaand waterbekken van niet te grote omvang (maar groter dan een poel)
waterpoel	ondiep, stilstaand waterbekken van geringe omvang
waterstroom	stromend water; rivier
waterval	omstandigheid dat of plaats waar stromend water van een hoogte of helling naar beneden valt
waterweg	weg te water, rivier of kanaal, geschikt voor scheepvaart
wereldzee	1 oceaan
wetering	1 water, stroom 2 beekje 3 gegraven water, groter dan een sloot
zee	1 uitgestrektheid zout water die het grootste deel van de oppervlakte van de aarde bedekt, syn. oceaan 2 elk gedeelte van de onder 1 genoemde uitgestrektheid, dat een eigen naam heeft
zeearm	lange smalle golf of inham van een zee
zeeboezem	baai[4]
zeeëngte	straat, nauwe doorgang van de zee tussen twee kusten
zeestraat	zeeëngte
zijrivier	rivier die uitmondt in een hoofdrivier
zijtak	2 zijkanaal; rivierarm

A.4 Italian Definitions

Bron:	**Nuovo Dizzionario Garzanti**
acqua	3 raccolta di acqua distesa di acqua
affluente	torrente o fiume che immette le sue acque in altro fiume maggiore
baia[1]	insenatura marina o lacustre larga al centro e stretta all'imboccatura
bassofondo	1 zona poco profonda di mare, pericolosa per la navigazione
bocca	3 breve passagio di mare tra due terre
braccio di fiume	ramo laterale
cala[1]	insenatura poco addentrata nella terraferma, adatta per l'approdo di piccole imbarcazioni

canale	1 sede artificiale di scorrimento di acque usate per l'irrigazione, la navigazione, l'industria, ecc. 2 zona di mare, per lo più stretta, compresa tra due terre opposte e vicine
cascata	brusca caduto d'acqua corrente causata da un dislivello
cateratta	2 ripida pendenza del letto di un fiume, che provoca un forte aumento di velocità nelle acque correnti
delta	2 pianatura alluvionale a forma di ventaglio che si osserva alla foce di un fiume
emissario	1 fiume che raccoglie e scarica le acque di un lago
estuario	foce di diume allargata a imbuto, in cui penetrano le maree
fiordo	insenatura profondo e angusta, propria di coste alte sottoposte a un'intensa glaciazione
fiume	1 corso d'acqua a corrente perenne e a regime pressoché costante
foce	zona di sbocco di un fiume nel mare o in un lago
fossato	lunga fossa, per lo più con acqua; piccolo corso d'acqua
fosso	a solco naturale o artificiale usato per lo scolo o la distribuzione dell'acqua
golfo	ampia e profonda insenatura della costa
guado[1]	il punto in cui un corso d'acqua si può guadare
idrovia	via di communicazione costituita da fiume e canali navigabile
immissario	corso d'acqua che si versa in un lago o in un altro bacino
insenatura	piccola rientranza della costa del mare o di un lago, o delle sponde di u fiume
lago	1 depressione del suolo occupata da acqua per lo piú dolce, non in diretta comunicazione col mare
laguna	tratto di mare basso separato dal mare aperto da una lingua di sabbia
mare	1 il complesso delle acque salate che comprono gran parte della superficie terrestre; quella parte di esse che é circondata da terre
oceano	vasta distessa di acqua salata che separa i continenti
pantano	1 terreno fangoso e coperto d'acque stagnanti
pozza	piccola cavità, depressione del terreno piena d'acqua
pozzanghera	pozza d'acqua fangosa
rigagnolo	piccolo corso d'acqua, spec. come quelli che scorrono ai lati delle strade quando piove
rivo	piccolo corso d'acqua o di alto liquido, sin. ruscello
ruscello	piccolo corso d'acqua, sin. rivo
secca	1 zona in cui il fondale marino, essendo poco profondo rispetto alla superficie dell'acqua, ostacola la navigazione
seno (di mare)	3 piccola insenatura
sorgente	1 getto d'acqua che scaturisce dal sottosuolo; il punto in cui l'acqua sgorga

stagno[2]	specchio d'acqua stagnante, poco profondo e poco esteso
stretto (di mare)	1 braccio di mare tra due terre, che congiunge due mari
torrente	breve corso d'acqua montano con forte pendenza e velocità e soggetto ad eccessi di magra e di piena
tributario	3 si dice di fiume che versa le sue acque in altro fiume o lago, sin. affluente
vasca	grande recipiente in pietra, in ferro o in altri materiali, spesso affondato nel terreno, per raccogliere acqua o altri liquidi

A.5 German Definitions

Bron:	**Duden Deutsches Universalwörterbuch**
Abflußgraben	Graben, in dem etw. abfließen kan
Abflußrinne	vgl. Abflußgraben
Bach	**1.** kleiner Natürlicher Wasserlauf von geringer Tiefe u. Breite **2.** Rinnsal, das sich aus abfließendem Regenwasser, Schmutwasser o.ä. gebildet hat
Arm	**2.** armartiger, armförmiger [Körper]teil; schmaler, seitlich abstehender, abzweigender Teil
Bächlein	Vkl. zu ↑ Bach
Bai	Meeresbucht, Meerbusen
Barre	Sandbank, Untiefe, bes. an der Mündung eines Flusses
Bergsee	See in den Bergen
Bodden	flacher Strandsee, flache Meeresbucht
Bucht	**1** [bogenartig] in das Land hineinragender Teil eines Meeres od. Binnengewässers
Delta	aus Schwemmland bestehendes, von den Mündungsarmen eines Flusses durchzogenes, deltaförmiges Gebiet im Bereich einer Flußmündung
Deltamündung	mehrarmiges Flußmündung, in deren Bereich sich ein deltaförmiges Schwemmland gebildet hat
Fjord	[an einer Steilküste] tief ins Landinnere hineinreichender, langgestreckter Meeresarm
Flach	(Seemannsspr.): nicht tiefe Stelle im Meer od. Fluß; Untiefe (1)
Fluß	1. größerer natürlicher Wasserlauf
Flußmündung	Mündung eines Flusses (1)
Flußarm	Arm (2) eines Flusses
Furt	seichte Stelle eines Flusses, die das Überqueren gestattet
Gebirgsee	See im Gebirge
Gewässer	größere natürliche Ansammlung von Wasser
Graben	1. [für einen bestimmten Zweck ausgehobene] längere, schmale Vertiefung im Erdreich

Gracht	schiffbarer Kanal in niederländischen Städten
Haff	durch eine Nehrung od. Inseln vom offenen Meer abgetrenntes Gewässer an einer Flachküste
Kanal	1. künstlicher schiffbarer Wasserlauf als Verbindung zwischen Meeren, Flüssen, Seen 2. offener Wasserlauf od. unterirdisch geführte Roherleitung für Abwässer, Bewässerung od. Entwässerung
[2]**Lache**	kleinere Ansammulung von Flüssichkeit, bes. von Wasser, die sich auf eine Fläche, en einer flachen Vertiefung gebildet hat
Lagune	vom offenen Meer durch einen Streifen Land, Riffe, o.ä. abgetrenntes Wasser
Maar	meist mit Wasser gefüllte, kraterförmige Senke vulkanischen Ursprungs
Meer	sich weitlich ausdehende, das Festland umgebene Wassermassen, die einen großen Teil der Erdoberfläche bedecken
Meeresarm	einem Fjord ähnliche, schmale, langgestreckte Bucht
Meeresbucht	bogenartig in das Land hineinragender Teil eines Meeres
Meeresstraße	**1** Meerenge **2** Seeschiffahrtstraße
Mündung	**1** Stelle, an der ein Fluß o.ä. mündet
Nebenfluß	Fluß, der in einen anderen Fluß mündet
Ozean	große zusammenhängende Wasserfläche zwischen den Kontinenten
Pfuhl	1. kleiner Teich, Ansammlung von schumtzigem, fauligem Wasser
Rigole	tiefe Rinne, kleiner Graben zur Entwässerung
Rinne	**1** schmale, langgestreckte Vertiefung im Boden, durch die Wasser fließt od. fließen kann
Ringgraben	vgl. Ringmauer
Rinnsal	1. sehr kleines, sacht fließendes Gewässer
[1]**See**	größere Ansammlung von Wasser in einer Bodenvertiefung des Festlandes; stehendes Binnengewässer
Seestraße	über das Meer führende Route, von Schiffen befahrene Strecke
Seitenarm	Arm (2)
Stadtgraben	der Befestigung einer Stadt dienender, um die Stadtmauer führender Graben
Staubecken	Becken für gestautes Wasser
Stausee	durch eines Flusses eintstandener See
Strom	1. großer, breiter (meist ins Meer mündender) Fluß
Teich	kleineres stehendes Gewässer, kleiner See
Trichtermündung	trichterförmige Mündung (eines Flusses)
Tümpel	Ansammlung von Wasser in einer kleineren Senke, Vertiefung im Boden
Untiefe	**1** flache, seichte Stelle in einem Gewässer

Wasser	2. Gewässer
Wasserfall	über eine od. mehrere Stufen senkrecht abstürzendes Wasser eines Flusses
Wasserloch	Erloch, in der sich Wasser angesammelt hat
Wasserstraße	von Schiffen befahbares Gewässer als Verkehrsweg
Weiher	kleiner See
Weltmeer	Ozean
Zufluß	2. in ein anderes Gewässer fließender Bach, Fluß

A.6 French Definitions

Source:	**Le Nouveau Petit Robert**
affluent	Cours d'eau qui se jette dans un autre
anse	**2** petite baie peu profonde
1. **baie**	échancrure d'une côte plus ou moins ouverte sur le large (en générale plus petit qu'un golfe)
bas-fond	**1** Partie du fond de la mer, d'un fleuve, où l'eau est peu profonde par rapport aux points voisins mais où la navigation est practicable (à la difference du haut-fond)
bassin	**2** Construction, ordinairement en pierre, destinée à recevoir de l'eau **3** Enceinte, partie d'un port, fluvial ou maritime, délimitée par des ouvrages (jetées, etc.) et dans laquelle les navires sont à flot **4** *Le bassin d'un fleuve:* le territooire arrosé par ce fleuve et ses affluents
bief	**1** Portion d'un cours d'eau entre deux chutes, d'un canal de navigation entre deux ćluses **2** Canal de dérivation qui conduit les eaux d'un cours d'eau vers une machine hydraulique
bisse	*Région.* (Suisse) Long canal d'irrigation conduisant l'eau des montagnes au sommet d'un terrain cultivé
bras	**5** Division d'un cours d'eau que partage des îles
bras de mer	détroit, passage
calanque	Crique étroit et allongée, bordée des rochers abruptes (spécialement en Méditerranée)
canal	**I. 1** Lit ou partie d'un cours d'eau **2** Cours d'eau artificiel **3** Bras de mer
cascade	**1** Chute d'eau; succession de chutes d'eau
cascatelle	*Littér.* Petite cascade
cataracte	Chute des eaux (d'un grand cours d'eau)

chenal	**1**Passage ouvert à la navigation entre un port, une rivière ou un étang et la mer, entre des rochers, des îles, dans le lit d'un fleuve **2** Courant d'eau établi pour le service d'une usine, le fonctionnement d'un moulin. **3** *Géol.* Sillon allongé dans une surface recouverte périodiquement ou constamment par les eaux
chute d'eau	Déplacement vertical d'une masse d'eau produit par la différence de niveau entre deux parties consécutives d'un cours d'eau
cours	**1** Ecoulement continy de l'eau (des fleuves, des rivières, des ruisseaux) **2** COURS D'EAU eau courante concentrée dans un chenal
crique	Enfoncement du rivage où les petits bâtiments peuvent se mettre à l'abri.
delta	Dépôt d'alluvions émergeant à l'embouchure d'un fleuve et la divisant en bras de plus en plus ramifiés.
détroit	**1** Bras de mer entre deux terres rapprochées et qui fait communiquer deux étendues marines.
douve	**I. 1.** Fossé rempli d'eau, autour d'un château, servant généralement à la défense. **2.** Large fossé précédé d'une barrière, dans un parcours de steeple-chase.
embouchure	**1** Ouverture extérieure. *Spécialt* Ouverture par laquelle un cours d'eau se jette dans une mer ou un lac.
étang	Étendue d'eau reposant dans une cuvette à fond imperméable et généralement moins vaste, moins profonde que le lac.
fjord	Ancienne vallée glaciaire envahie par les eaux marines durant la déglaciation, caractéristique des côtes scandinaves et écossaises.
flaque	Petite nappe de liquide stagnant.
fleuve	**1** *Cour.* Grande rivière (remarquable par le nombre de ses affluents, l'importance de son débit, la longueur de son cours); spécialt lorsqu'elle aboutit à la mer. ▷ *Géogr.* Cours d'eau (même petit) aboutissant à la mer
fondrière	Affaissement, trou plein d'eau ou de boue dans un chemin défoncé
fossé	**1** Fosse creusée en long dans le sol et servant à l'écoulement des eaux, à la séparation des terrains *Fortif.* Tranchée entourant un ouvrage fortifié et servant à la défense.
gave	Cours d'eau, torrent pyrénéen.
golfe	Vaste bassin en cul-de-sac plus ou moins largement ouvert, que forme la mer dans son avancée à l'intérieur des terres
gué	Endroit d'une rivière o le niveau de l'eau est assez bas pour qu'on puisse traverser à pied.
lac	Grande nappe naturelle d'eau douce ou (plus rarement) salée, à l'intérieur des terres

lagune	tendue d'eau de mer, comprise entre la terre ferme et un cordon littoral (lido) généralement percé de passes (graus).
océan	**1** Vaste étendue d'eau salée qui couvre une grande partie de la surface du globe terrestre.
lagon	**1** Petit lac d'eau salée, lagune peu profonde entre la terre et un récif corallien, par les brèches duquel pénètre la marée. **2** Lagune centrale d'un atoll.
marais	**1** Nappe d'eau stagnante généralement peu profonde recouvrant un terrain partiellement envahi par la végétation
mare	**1** Petite nappe d'eau peu profonde qui stagne
mer	**1** Vaste étendue d'eau salée qui couvre une grande partie de la surface du globe. **2** Bassin océanique, plus ou moins isolé, de dimensions limitées.
nappe	*Géol. Nappe (d'eau)* : eau occupant une dépression fermée
oued	Rivière d'Afrique du Nord, cours d'eau temporaire dans les régions arides.
passe	**II. 2.** *Géogr., mar.* Passage étroit ouvert à la navigation.
pertuis	**2** *Mod. Techn.* Ouverture qui permet de retenir l'eau d'une écluse ou de la laisser passer. ◇ *Géogr.* tranglement d'un fleuve.
piscine	**2** *Cour.* Grand bassin de natation, et ensemble des installations qui l'entourent.
rade	Bassin naturel de vastes dimensions, ayant issue vers la mer et dans lequel les navires peuvent trouver un bon mouillage.
réservoir	**1** Bassin où un liquide peut être gardé en réserve.
rivière	**I. 1.** Cours d'eau naturel de moyenne importance. **2** Sport Fossé rempli d'eau que doit sauter le cheval (steeple-chase) ou le coureur (steeple).
ruisseau	**1** Petit cours d'eau, affluent d'une rivière, d'un lac, d'un étang
seguia	Canal d'irrigation, en Afrique du Nord.
torrent	**1** Cours d'eau à forte pente, à rives encaissées, à débit rapide et irrégulier.

A.7 Russian Definitions

Source:	**Толковый Словарь Русского Языка**
бассейн	**1** Искусственный водоём, сооружённый для плавания, купания, в декоративных целях **2** Совокупность притоков реки, озера, а также площадь стока поверхностных и подземных вод в водоём
брод	Мелкое место в реке, озере, удобное для перехода
бухта	Небольшой глубокий залив
водоём	место скопления или хранения воды (озеро, бассейн, пруд, водохранилище)

водопад	Стремительно падающий с высоты поток воды
водохранилище	Водоём, в котором скапливается и сохраняется вода
залив	Часть водного пространства, вдавшаяся в сушу
канава	Неглубокий и неширокий ров
канал	**1** Искусственное русло, наполненное водой
лагуна	**1** Морской залив, отделённый от моря песчаной косой
	2 Внутренний водоём коралловых островов, а также участок моря между коралловым рифом и берегом
лиман	Залив, образованный морем в низовьях реки, а также солёное озеро вблизи моря, обычно богатое целебными грязями
лужа	Небольшое углубление на почве, наполненное дождевой или подпочвенной водой
	2 Пролитая на поверхность жидкость
мелководье	Низкий уровень воды в реке, водоёме
мель	Мелководное место в реке, водоёме
море	Часть океана – большое водное пространство с горько солёной водой
озеро	Замкнутый в берегах большой естественный водоём
океан	Весь водный покров Земли, окружающий материки и острова
поток	Стремительно текущая водная масса
приток	**3** Река, впадаюшая в другую реку или в озеро
пролив	Узкое водное пространство, разделяющее участки суши и соединяющее смежные бассейны или их части
пруд	Водоём в естественном или выкопанном углублении, а также запруженное место в реке
река	Постоянный водный поток значительных размеров с естественным течением по руслу от истока до устья
речка	Небольшая река
ров	Длинное, с высокими откосами углубление в земле
рукав	**2** Ответвление от главного русла реки, гл. обр. в её устье
русло	Углубление в грунте, по к-рому течёт водный поток
ручей	Водный поток, текущий струёй
устье	**1** Место впадения реки (в море, озеро или другую реку)
фьорд	Узкий, глубоко вдавшийся в берег морской залив со скалистыми берегами

Appendix B

Abbreviations and Notations

B.1 Dictionaries Referred to in this Thesis

COD Concise Oxford Dictionary of Current English. First edited by H. W. Fowler & F. G. Fowler, 1911. Eighth edition, edited by R. E. Allen. Oxford: Clarendon Press. 1990.

Collegiate Merriam-Webster's Collegiate Dictionary. Tenth edition, edited by Frederic C. Mish. Springfield: Merriam-Webster, Inc. 1993.

Duden Duden Deutsches Universalwörterbuch. Second edition, edited by Günther Drowdowski. Mannheim: Dudenverlag. 1989.

Garzanti Il Nuovo Dizionario Italiano Garzanti. First edition, edited by Schiannini. Milano: Garzanti Editore, 1984.

GVD van Dale Groot Woordenboek der Nederlandsche Taal. First edited by I.M. Calisch & N. S. Calisch, 1864. Twelth edition, edited by prof. dr. G. Geerts & dr. H. Heestermans. Utrecht: van Dale Lexicografie. 1992.

Hachette Dictionnaire Hachette de la Langue Française. Paris: Hachette.

Larousse Nouveau Petit Larousse. Third edition. Paris: Librairie Larousse. 1970.

LDOCE Longman Dictionary of Contemporary English. Second edition. Essex: Longman. 1987.

OED Oxford English Dictionary. First published in 1928. Second edition, edited by John Simpson and Edmund Weiner. Oxford: Oxford University Press. 1989.

Oxford Hachette The Oxford-Hachette French Dictionary - Le Grand Dictionnaire Hachette-Oxford. First edited by M.-H. Corréard & V. Grundy,

1994. Third edition, edited by J.-B. Ormal-Grenon & N. Pomier. Paris: Hachette Livre. 2001.

Oxford Zanichelli Il Ragazzini/Biagi Concise Dizionario Inglese Italiano - Italian English Dictionary. Third edition, edited by G. Ragazzini & A. Biagi. Bologna: Grafica Editoriale Printing. 2001.

Ozhegov Толковый Словарь Русского Языка. First edited by Х. Ю. Шведова, 1968. Fourth edition, edited by Ц. И. Ожегов & Х. Ю. Шведова. Moscow, Sovietskaya Entsiclopedia. 1999.

Petit Robert Le Nouveau Petit Robert, Dictionnaire de la Langue Franąise. First edited by Paul Robert, 1967. Second edition, edited by Josette Rey-Debove & Alain Rey. Paris: Dictionnaires le Robert. 1993.

VDNE van Dale Handwoordenboek Nederlands - Engels. First edition, edited by dr. M. Hannay. Utrecht: van Dale Lexicografie. 1988.

VDEN van Dale Handwoordenboek Engels - Nederlands. First edition, edited by dr. M. Hannay. Utrecht: van Dale Lexicografie. 1988.

VDFN van Dale Handwoordenboek Frans - Nederlands. First edition, edited by dr. P. Bogaard. Utrecht: van Dale Lexicografie. 1988.

VDIN van Dale Handwoordenboek Italiaans - Nederland / Zanichelli Dizionario Italiano - Neerlandese. First edition, edited by prof. dr. V. Lo Cascio. Utrecht: van Dale Lexicografie. 2001.

Webster Mirriam-Webster 3rd New International Dictionary, unabridged. First edition, edited by Philip Babcock Grove and the Merriam-Webster editorial staff. Springfield: Merriam-Webster, Inc. 1961.

B.2 IPA Pronuncation Rules

This is not a complete list of all IPA symbols, but the list of all possible unclear IPA symbols used in this thesis.

ɑ	script a	f**a**ther
ɐ	turned a	*English* b**o**dy
ɒ	turned script a	s**o**rry
a	a	*Italian* p**a**sta
æ	a-e	st**a**llion
ɛ	epsilon	**e**nd
e	e	**ei**ght
g	g	**g**et
ɪ	iota	br**i**dge
i	i	fill**y**
j	j	**y**our
ə	schwa	tow**e**l
ʏ	small caps y	*German* Gl**ü**ck
ŋ	eng	ha**ng**
ɔ	open o	c**au**ght
o	o	*French* b**eau**
ʃ	esh	**sh**ower
θ	theta	**th**in
ɥ	turned h	*French* r**ue**
ʊ	upsilon	p**u**ll
u	u	**oo**ze
ʌ	turned v	h**u**rry
œ	o-e	*French* h**eu**re
x	x	*Scots* lo**ch**
y	y	**y**our
ʤ	d-yogh	**j**og
ː	-	lengthen preceding vowel
ˈ	-	primary stress
ˌ	-	secundary stress

B.3 Lexical Functions

Function	Description	Example
A_0	adjective	A_0(law)=legal
A_i	typical modifier for ith actant	A_i(surprise)=surprised
$Able_i$	adjective for capability of ith actant	$Able_i$(read)=literate
Adv_0	adverb	Adv_0(honest)=honestly
Adv_i	adverbial for ith actant	Adv_i(dismay)=in dismay
Anti	antonym	Anti(like)=dislike
Bon	"good" (expression of praise)	Bon(advice)=sound
Cap	leader/chief	Cap(school)=head
Caus	cause	Caus(rise)=raise
CausPredMinus	cause to decrease	CausPredMinus(price)=drop
CausPredPlus	cause to increase	CausPredPlus(price)=increase
Centr	centre/middle	Centr(problem)=crux
Cont	continue	$ContOper_1$(influence)=maintain
Contr	contrastive term	Contr(heaven)=earth
$Conv_{ij}$	converse term	$Conv_{21}$(more)=less
Culm	culmination	Culm(anger)=paroxysm
Degrad	degrade, get worse	Degrad(milk)=go/turn sour
Equip	team, crew	Equip(hospital)=staff
Excess	function excessively	Excess(heart)=palpitate
$Fact_{0,1,2}$	be realized	$Fact_0$(dream)=come true
Figur	standard metaphor	Figur(smoke)=cloud
Fin	cease, stop	$FinOper_1$(influence)=lose
$Func_{0,1,2}$	nearly empty verb (keyword=subject)	$Func_0$(silence)=reign
Gener	superordinate	Gener(anger)=feeling
Germ	germ, core	Germ(evil)=root
Imper	order, command	Imper(silence)=shut up!
Incep	begin	$IncepFunc_0$(war)=break out
IncepPredMinus	start to decrease	IncepPredMinus(price)=fall
IncepPredPlus	start to increase	IncepPredPlus(price)=skyrocket
Instr	typical preposition (=with the help of)	Instr(car)=by
Involv	keyword = subject	Involv(smell)=fill [the room]
$Labor_{ij}$	nearly empty verb; ith actant = subject; jth actant = direct object	$Labor_{12}$(consideration)=take into

Liqu	liquidate, delete	Liqu(disease)=eradicate
$\mathrm{Loc}_{ab/ad/in}$	locative prepositions	Loc_{in}(list)=on
Magn	intensifier	Magn(bachelor)=confirmed
Manif	be manifest	Caus_1Manif(opinion)=express
Minus	less	IncepPredMinus(price)=fall
Mult	regular group/set	Mult(dog)=pack
Nocer	damage, attack	Nocer(mosquito)=bite
Obstr	function with difficulty	Obstr(voice)=falter
$\mathrm{Oper}_{1,2}$	nearly empty (support) verb (keyword=subject)	Oper_1(attention)=pay
Pejor	worse	CausPredPejor(prospect)=darken
Perf	perfective (completed action)	S_1Perf(marry)=spouse
Perm	permit	$\mathrm{Perm}_1\mathrm{Fact}_0$(passion)=succumb to
Plus	more	IncepPredPlus(price)=skyrocket
Pos_i	positive evaluation of ith actant	Pos_2(opinion)=favorable
Pred	predicate (= to be)	Pred(actor)=act
Prepar	prepare	$\mathrm{PreparFact}_0$(rifle)=load
Propt	typical preposition (= because of)	Propt(fear)=for
Prox	on the verge of	$\mathrm{ProxFunc}_0$(storm)=approach
Qual_i	Able i + highly probable	Qual_1(device)=deceitful
$\mathrm{Real}_{1,2,\ldots}$	satisfy the requirements of	Real_1(promise)=keep
Result	result of an event	Result(learn)=know
S_0	noun	S_0(die)=death
S_i	typical noun for ith actant	S_1(murder)=murderer
S_{instr}	typical instrument	S_{instr}(paint)=brush
S_{loc}	typical place	S_{loc}(lion)=den
S_{med}	typical means	S_{med}(write)=ink
S_{mod}	typical mode	S_{mod}(write)=handwriting
S_{res}	typical result	S_{res}(copy v)=copy n
Sing	regular "portion"	Sing(rice)=grain
Son	typical sound	Son(elephant)=trumpet
Sympt	physical symptoms	Degrad(speech) + Sympt_{23}(surprise)=be speechless
Syn	synonym	Syn(help)=aid
V_0	verb	V_0(advice)=advise
Ver	as it should be	Ver(excuse)=legitimate

B.4 Notational Conventions used in this Thesis

Name	Notation	Description
Interlingual Meaning	SMALLCAPS	An interlingual meaning is a word-sense expressed by at least one of the lexemes of at least one of the languages: STALLION is a meaning of the English lexeme *stallion*
Definitional Attribute	**boldface**	A definitional attribute is a feature of the interlingual meaning it relates to; a differentiam specificam of the related lexeme: **male** is a feature an interlingual meaning can have to indicate that the lexeme expressing that interlingual meaning has *male* in its lexical definition.
Formal Concept	**BOLD SMALLCAPS**	Formal Concepts have in general no names, but the smallest common concept of an interlingual meaning is indicated in bold smallcaps: **COLT** $:= \langle \text{COLT}'', \text{COLT}' \rangle$
Lexeme	*slanted*	The lexeme is the actual lexical entry in the dictionary, the headword of which is the citation-form.
String	`courier`	A string or orthographic word is a sequence of letters.
Word-form	sans-serif	A word-form is an abstract representation of a word, cutting across spelling and pronunciation; it consists of a number of spelling-cum-pronunciations, a word-class and possibly a gender.
Phonological Word	/fəˈnetɪk/	A phonological word is a spoken word, identified with a prototypical pronunciation, represented in IPA between slashes.

Nederlandse Samenvatting

Men neemt algemeen aan dat er in de orde van vijf- tot zesduizend talen zijn. Afgezien van het Engels, Frans of het Spaans, bestaat er voor veel talenparen $\langle X, Y \rangle$ niet een woordenboek $X \rightarrow Y$ of $Y \rightarrow X$. Men moet het dan meestal doen met woordenboeken $X \rightarrow$ Engels/Frans/Spaans en Engels/Frans/Spaans $\rightarrow Y$. Toch is er een maatschappelijke behoefte aan vertaalwoordenboeken die de leden van een paar direct met elkaar in een vertaalrelatie brengen zonder de tussenkomst van een klein aantal West-Europese talen met een koloniaal verleden. Ook op theoretische gronden is een dergelijke behoefte te verdedigen.

Het maken van een kwalitatief goed woordenboek vergt veel tijd, en daar er uit de vijf- tot zesduizend talen zo'n 25 tot 30 miljoen talenparen zijn, is het van belang een database te hebben, op grond waarvan directe vertalingen tussen talen mogelijk worden gemaakt. Het proefschrift brengt enkele problemen in kaart die zich bij zo'n onderneming voordoen, tracht enkele daarvan op te lossen en van andere aan te tonen dat de weg niet begaanbaar is.

Een bekend probleem is dat woorden uit verschillende talen moeilijk op elkaar te passen zijn: woorden in verschillende talen hebben vaak niet hetzelfde bereik aan betekenissen, niet alle woorden uit de ene taal hebben een equivalent in een andere, etc. In dit proefschrift geef ik een aanzet tot de opzet van een database waarin een groot deel van deze problemen opgelost wordt. Cruciaal in deze opzet is de structurering van de tussentaal, waarmee in de database niet-corresponderende betekenissen toch op gestructureerde wijze aan elkaar gerelateerd kunnen worden. De structuur van deze tussentaal wordt geleverd door een logisch raamwerk, onder de naam Formele Begripsanalyse. Met deze opzet kan onder meer voor woorden waarvoor geen directe vertaling is in de doeltaal toch een omschrijvende vertaling gegenereerd worden. Daarmee wordt het werk van een lexicograaf die een vertaalwoordenboek voor een talenpaar moet maken vergemakkelijkt.

De wijze waarop dit proefschrift is opgebouwd is als volgt. In hoofdstuk 1 van dit proefschrift wordt een beschouwing gegeven op de vraag aan welke eisen een lexicaal gegevensbestand moet voldoen om in staat te zijn niet-corresponderende betekenissen op gestructureerde wijze aan elkaar te

verbinden. Hierbij wordt een aantal bestaande lexicale gegevensbestanden onder de loep genomen om te kijken om welke reden dit in die systemen op dit moment niet mogelijk is.

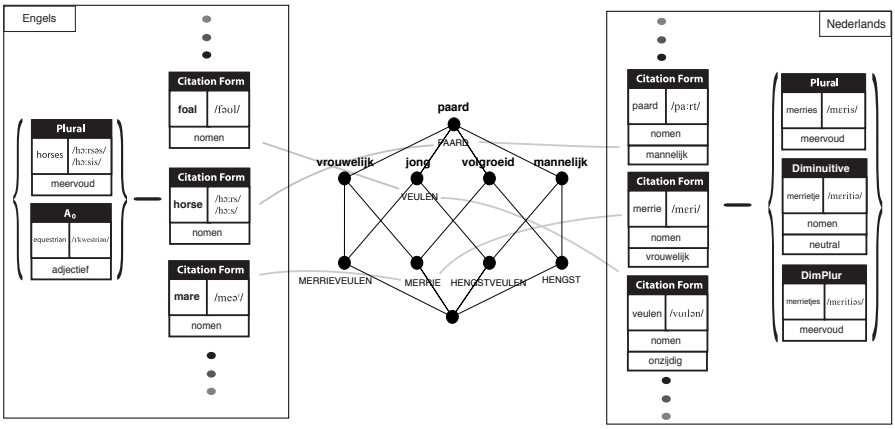
In hoofdstuk 2 wordt de opzet van het SIM*u*LLDA-systeem geschetst. In deze opzet worden de verschillende talen in het lexicale gegevensbestand aan elkaar gekoppeld door middel van een tussentaal (interlingua). En zoals gezegd is de structuur van deze tussentaal het hart van het SIM*u*LLDA-systeem. De tussentaal is een gestructureerde eenheid, bestaande uit een traliestructuur. De knopen van deze tralie bestaan uit paren van taalonafhankelijke betekenissen en eigenschappen van deze betekenissen die *definitionele attributen* genoemd worden.

De taalonafhankelijke betekenissen zijn alle betekenissen die door een van de woorden (of eigenlijk: lexemen) uit een van de talen worden uitgedrukt. Dus ieder lexeem uit iedere taal drukt een of meer betekenissen uit. Het omgekeerde is echter niet waar: niet iedere betekenis is in elke taal gelexicaliseerd. Stel bijvoorbeeld dat er in het Italiaans een woord is voor *wegen die naar Rome leiden*, en dat er geen enkele andere taal is met een dergelijk woord. Dan is deze betekenis nog steeds een taalonafhankelijke betekenis, echter een waarvoor alleen in het Italiaans een concreet woord bestaat. De andere talen hebben dan wat heet een *lexicale leemte* ten aanzien van deze betekenis, c.q. dit Italiaanse woord.

De definitionele attributen zijn de eigenschappen die de taalonafhankelijke betekenissen vastleggen. Deze definitionele attributen vinden hun herkomst in de *differentiae specificae* in monolinguale woordenboeken. Om een voorbeeld te geven: het woord hengstveulen is gedefinieerd in een woordenboek als een *jong mannelijk paard*. In deze definitie worden **jong** en **mannelijk** opgevoerd als kenmerkende eigenschappen van de betekenis van het woord hengstveulen. Het zijn deze kenmerken die gelden als definitionele attributen in SIM*u*LLDA. De *genus proximum* in deze definitie (paard) duidt geen definitioneel attribuut aan, maar verwijst naar een andere betekenis in het woordenboek, waar weer nieuwe definitionele attributen bijhoren.

Als we even afzien van de doorverwijzing naar paard en **paard** wel degelijk beschouwen als een definitioneel attribuut, krijgen we een structuur als weergegeven in Figuur 1.

Definitionele attributen zijn, zoals gezegd, eigenschappen die de betekenissen in SIM*u*LLDA vastleggen. Maar daar deze betekenissen taalonafhankelijk zijn, kunnen deze attributen zelf nooit taalspecifiek zijn. Derhalve zijn ook definitionele attributen taalonafhankelijke elementen van de gestructureerde tussentaal, die in elk van de talen in het lexicale gegevensbestand gelexicaliseerd kunnen worden. Daarbij is er ook een lexicalisatie voor een definitioneel attribuut als dit attribuut in de gegeven taal geen rol speelt.



Figuur 1: Opzet van het SIMuLLDA-Systeem

Om terug te keren op het eerder genoemde voorbeeld: het deel *die naar Rome leiden* zal een Nederlandse lexicalisatie van een definitioneel attribuut zijn (**naar_Rome_voerend**). En dan specifiek een definitioneel attribuut dat bij de bepaling van geen enkel Nederlands woord een rol speelt.

De opzet in Figuur 1 maakt het mogelijk vertalingen voor woorden te geven: het lexeem horse is gekoppeld aan de betekenis HORSE en de betekenis HORSE is weer gekoppeld aan het Nederlandse woord paard. Dus paard en horse zijn rechtstreekse vertalingen of 'vertalingssynonymen' van elkaar.

Door de structuur van de tussentaal wordt het echter ook mogelijk omschrijvende vertalingen te geven voor woorden waarvoor geen directe vertaling bestaat. Een voorbeeld aan de hand van de tussentaal in figuur 1: het Nederlandse woord hengstveulen kent geen rechtstreekse vertaling in het Frans: er is wel een woord voor merrieveulen (pouliche) en een algemener woord voor veulen (poulain), maar er is geen woord voor hengstveulen als zodanig.

Door de plaatsing van de betekenis HENGSTVEULEN in de tralie kunnen we echter wel van alles zeggen over deze betekenis. Allereerst hangt de knoop waarbij deze betekenis hoort onder de knoop van de betekenis VEULEN, en is de betekenis VEULEN wel gelexicaliseerd in het Frans: poulain. Dus vanuit de tralie kunnen we stellen dat poulain een redelijke, zij het iets te algemene vertaling is voor hengstveulen. Dat deze vertaling te algemeen is komt doordat HENGSTVEULEN meer definitionele attributen heeft dan VEULEN: het heeft een *definitioneel surplus*. Dit definitionele surplus bestaat uit precies één definitioneel attribuut: **mannelijk**. Dus wat mist in de poulain-vertaling is **mannelijk**, wat in het Frans kan worden uitgedrukt met *mâle*. De complete betekenis van hengstveulen in het Frans is de combinatie van deze twee: *poulain mâle*.

Gegeven de manier waarop lexicale leemten worden opgevuld is de notie van *differentiae specificae* in SIM*u*LLDA geheel op het niveau van de tussentaal vastgelegd: HENGSTVEULEN = VEULEN + **mannelijk**. Het is ook mogelijk de rechterzijde van deze vergelijking weer terug in het Nederlands te vertalen. Dit levert een lexicale definitie op: **hengstveulen** - *mannelijk veulen*. Dus met de SIM*u*LLDA-opzet is het mogelijk zowel lexicale definities te genereren als vertalingen voor tweetalige woordenboeken, ook in het geval er geen vertalingssynonym bestaat.

Tenslotte worden in hoofdstuk 2 ook nog enige logische eigenschappen besproken van het systeem dat ten grondslag ligt aan SIM*u*LLDA: Formele Concept Analyse. FCA zorgt ervoor dat de relatie tussen definitionele attributen en taalonafhankelijke betekenissen de traliestructuur opleveren die is weergegeven in Figuur 1. Als onderdeel hiervan wordt ook een internet-applicatie geïntroduceerd die ten behoeve van dit proefschrift ontwikkeld is: JaLaBA, een programma dat de omzetting van tabellen naar de traliestructuur automatisch uitvoert, en een 3-dimensionale weergave van de tralie geeft. Deze applicatie is, net als alle andere aan dit proefschrift gerelateerde zaken, te vinden op de web-site van dit proefschrift: `http://maarten.janssenweb.net/simullda`.

In de hiervoor beschreven opzet van SIM*u*LLDA zit een aantal verborgen premissen: ieder woord drukt een vast aantal betekenissen uit, woorden uit verschillende talen kunnen dezelfde betekenis uitdrukken en betekenissen worden vastgelegd door middel van definitionele attributen. Al deze premissen zijn echter problematisch. Met name de laatste doet sterk denken aan de reeds lang verworpen stelling dat woordbetekenis terug te voeren is op een kleine set aangeboren semantische primitieven. Om desondanks de opzet te hanteren zoals hierboven beschreven is het daarom van groot belang precies aan te geven wat de status van de verschillende onderdelen van het systeem is: wat woorden, talen, betekenissen en definitionele attributen precies geacht worden te zijn. En met name ook welke claims niet worden gemaakt door het systeem.

Een diepgaande beschouwing van alle onderdelen wordt gegeven in hoofdstuk 3: talen worden gedefinieerd als willekeurige verzamelingen lexemen, waarbij lexemen de ingangen in een woordenboek zijn. Lexemen bestaan op hun beurt weer uit verzamelingen woordvormen, te weten de vervoegingen en verbuigingen van het woord. En de woordvormen zijn representaties van abstracte entiteiten, bestaande uit een uitspraak, een spelling, een woordklasse en evt. een geslacht. Lexemen zijn geen woorden in de standaardzin van het woord: lexemen kenmerken zich niet doordat ze door spaties gescheiden worden; lexemen kunnen ook uit meerdere woorden bestaan (bv. bij idiomatische uitdrukkingen) of kleinere delen omvatten (morfemen).

Bij de karakterisering van de betekenissen en definitionele attributen is het vooral van belang aan te geven wat deze niet zijn: betekenissen zijn

niet denotationeel van karakter. Dat wil zeggen, betekenissen zijn niet gelijk aan noch worden bepaald door de verzameling van objecten die onder het begrip vallen; betekenissen leggen ook niet vast welke objecten er precies onder vallen, noch stelt de betekenis je in staat van ieder object eenduidig vast te stellen of het onder dat begrip valt of niet. Daarnaast zijn de definitionele attributen die de betekenissen vastleggen geen zwaar fundamentele atomen zoals Katz & Fodor hebben voorgesteld: ze zijn niet aangeboren, er is niet een van God gegeven aantal attributen, en definitionele attributen leggen niet alles vast wat we doorgaans onder woordbetekenis laten vallen. Van veel begrippen weten we hoe het kenmerkende element ervan er uit ziet (bv. wat een typisch ontbijt is), maar dergelijke prototypen zijn niet taalonafhankelijk en worden ook niet vastgelegd door definitionele attributen. Definitionele attributen zijn niet meer en niet minder dan de differentiae specificae uit woordenboeken.

De opzet van dit proefschrift is het leveren van een werkbaar model dat het werk van lexicografen verlicht. Daarom is het van groot belang te laten zien dat de hierboven beschreven opzet in praktijk ook daadwerkelijk bruikbaar is. Hoewel de echte bruikbaarheid slechts blijkt als het systeem op grote schaal gebruikt zou worden, wordt in hoofdstuk 4 van dit proefschrift een beperkte empirische test beschreven: als we alle woorden voor wateren (meren, rivieren, plassen, baaien, etc.) in een zestal talen in oogschouw nemen, is het mogelijk deze met behulp van SIM*u*LLDA aan elkaar te koppelen. Dit levert dan daadwerkelijk de gewenste situatie waarbij voor ieder woord in iedere taal een vertaling in iedere taal gegenereerd kan worden. Er is een tweede, kleinere empirische test voor alle namen van zeilen op een vijf-master. Dit laatste voornamelijk om een vergelijking te trekken met de analyse die gegeven wordt in het proefschrift van Marc van Campenhoudt. Hoewel er zich bij deze empirische tests enige problemen voordoen, zijn deze problemen (op een na) met kleine aanpassingen van het systeem op te lossen.

Het enige probleem dat niet op te lossen is binnen het systeem is het volgende: een woordenboekdefinitie omschrijft de betekenis van een woord door (naast de differentiae specificae) te verwijzen naar het genus proximum. Dit genus is de *betekenis* van een van de andere woorden in het woordenboek. Nu komt het echter om verschillende oorzaken voor dat deze genusbetekenis zelf niet in het woordenboek staat; niet door een leemte, maar doordat het ondergespecificeerd is wat de genus betekenis precies is. De reden hiervoor is dat een basale aanname die impliciet achter woordenboeken schuil gaat, in feite incorrect is: woorden hebben geen welomschreven aantal onderscheiden betekenissen. Een goed voorbeeld hiervan is dat bv. het woord raam twee betekenissen heeft die onderscheidbaar maar toch ook hetzelfde zijn: het is een opening in de zijkant van een huis, maar ook het glas dat daarin besloten zit. Als we hiervoor twee betekenissen opnemen, moeten we ook van alle hyponymen van raam aannemen dat

ze ambigu zijn: een dakraam is een raam in deze beide betekenissen samen. De stelling is ook dat binnen een systeem waarin woordenboek definities serieus worden genomen, dit probleem niet opgelost kan worden.

In zijn basale opzet behelst het systeem slechts een beperkt gedeelte van de in woordenboeken aanwezige informatie: alleen de semantische definities en dan ook nog alleen van nomina. Om een volledige lexical gegevensbestand te zijn dienen de andere delen van woordenboekinformatie echter ook een plaats te krijgen in het systeem. Dit wordt ten dele in hoofdstuk 5 opgelost. In dat hoofdstuk wordt beschreven hoe labels, collocaties, voorbeeldzinnen en morfologische derivaties kunnen worden gemodelleerd in het systeem, deels gebruik makend van de lexicale functie uit de Meaning⇔Text Theory. Ook wordt kort besproken hoe het systeem zich verhoudt tot andere woordklassen dan nomina, zoals werkwoorden en adjectieven en worden enkele aspecten beschreven van een eventueel voor dit systeem te ontwikkelen toepassing.

Curriculum Vitae

I was born in Nijmegen on January 28, 1971. I attended the Nijmeegsche Scholen Gemeenschap (NSG) in Nijmegen, and obtained my Gymnasium diploma in 1989.

In 1989, I began my studies at Utrecht University, starting with Philosphy and Chemistry. In 1991, I obtained a 'propedeuse' in Chemistry, in 1993 another one in CKI (Cognitive Artificial Intelligence), and in 1995 also one in Algemene Letteren (General Linguistics). In 1997, I obtained my 'doctoraal examen' ($\geq$ Master's degree) in Cognitive Artificial Intelligence. My doctoraalscriptie ($\approx$ Master's thesis), called 'The Parsons Project', made a comparison between the Gruberian approach to events found in the PLUG theory by Henk J. Verkuyl, and the neo-Davidsonian approach to events found in 'Event Semantics' by Terrence Parsons.

In 1998, I started as a junior docent (junior lecturer) at the department of Philosophy in the Theoretical Philosophy and Logic Group, and also as a PhD-student at the Utrecht Institute of Linguistics OTS, in cooperation with Cognitive Artificial Intelligence. There I carried out the research that resulted in this dissertation.

Acknowledgements

It is my pleasure to take this opportunity to thank a number of people who contributed to the completion of this thesis.

First, I would like to thank my 'promotoren' (thesis advisors), Henk J. Verkuyl and Albert Visser. They have been supporting my scientific career since 1995, when I started my master's thesis on Event Semantics with both of them (and Michael Moortgat) as my supervisors. Not only have they been invaluable for the development of both my master's and my PhD thesis, but it has furthermore always been a great pleasure to work with them.

It should be said that the original subject of this thesis was rather different: it was a more ambitious attempt to give a model for word meaning that was at the same time philosophically plausible, psychologically realistic, and linguistically applicable. In this original set-up, I also had Renée van Hezewijk as a co-promotor and I would like to thank him for the time he invested in guiding me through the psychological literature, even though this aspect does not appear in the present thesis.

Jan van Eijck was the one who originally came up with the idea of investigating Formal Concept Analysis, albeit in a very different setting: his idea had to do with the previous subject of this thesis, and he suggested to model prototypes by means of FCA. As said in the thesis, the idea to apply FCA to dictionaries came from the thesis of Uta Priß, although I cannot remember how or why I came across her thesis.

I owe a lot of thanks to Vincent van Oostrom. He helped me with many logical puzzles, often explained new logical theories to me, and checked a number of proofs for this thesis. Furthermore, he often provided me with critical remarks, questions, suggestions, and last but not least many hours of pleasant conversation.

I would also like to thank Igor Mel'čuk who was kind enough to give me a great number of suggestions during the 1999 Batumi conference on Logic, Language, and Information, although I am ashamed to admit that I failed to follow up on most of them. Likewise, I would like to thank Willy Martin, Pepijn Visser, and Piek Vossen, Gilles Sérasset, Marc van Campenhoudt, and Paul Dekker, who provided helpful insights and comments at some point in the process.

I would like to thank the Utrecht Institute of Linguistics OTS and the

department of Philosophy (Opleidingsinstituut CKI) for providing me the opportunity to carry out my research. And I thank the members of the supervising committee (Willy Martin, Francisca de Jong, Keimpe Algra, Jan van Eijck and Michael Moortgat) for reading the manuscript.

On a more personal note, I would like to thank my paranimfen: Marije van Duijne Strobosch (my ex girlfriend, whom I would of course also like to thank for running out on me :-) and Christof Faassen, one of my oldest friends, who found time to help me despite the fact that he recently became a father. And also Peter Janssen and André van Schie who have been very helpful in the final stages of this thesis. It is probably due to their academic background that I decided to become a PhD student.

Of course, I also like to thank my parents for their patience and moral (and financial) support, my father for providing the illustration for the cover of this thesis, and my brother for offering to proof-read the text of this thesis (and nagging about my lack of progress all the time).

Finally, I would like to end with a list of names of people. People usually read this section to find their own name, so giving a long list of names should satisfy this wish for many. The people on this list may or may not have had anything to do with my thesis, and are a mixture of friends, family, and colleagues. The list is in alphabetic order, so the people I like best are mentioned first. I hope and assume that I forgot to put a great many people on this list. So here it goes: Anna Mlynarczyk, Annet Geerlings, Anouschka Busch, Astrid Porsius, Balazs Suranyi, Bart Drykoningen, Carolien Porsius, Daniëlle van den Burg, Dimitri Hendriks, Dirk Heylen, Ellen Geerlings, Ellen Gerrits, Elma Blom, Ernestine Elkenbracht, Esther Kraak, Floor Ypma, Frank Wijnen, Gera Heidema, Guus de Ruiter, Herman Hendriks, Igor Boguslavsky, Iris Mulders, Ivana Čače, Jan Bergstra, Janine Luijk, Jeldau Bollema, Johan Zuidema, Joost Joosten, Kate Lohoefener, Kees Vermeulen, Knud Boucher, Lev Beklemieshev, Lodewijk Regout, Lodewijk Wagenaar, Manuel Tjin-a-Ton, Menno Lievers, Mia Ypma, Michiel Fast, Mieke Stevens, Mies Dieben, Miki Logger, Nada Vasic, Nathalie Kröner, Nynke Visser, Oele Koornwinder, Olga Borik, Oscar Logger, Øystein Nilsen, Pascal Maas, Patrick Brandt, Paz Gonzalez, Pepijn Visser, Piroska Lendvai, Raffaella Bernardi, Raymond Cremers, Richard Moot, Rick Nouwen, Robbert van Rooy, Robert Voors, Roelof Bosch Rudolf Hoyng, Ruud Visser, Sanne Hartog, Saskia de Jong, Severine Rutgrink, Sharon Unsworth, Silke Hamann, Sylvia Huydecoper, Ted Hes, Willemijn Lindenhovius, Willemijn Vermaat, en Xavier van Buchem.

References

Al-Kasimi, Ali M. 1977. *Linguistics and Bilingual Dictionaries*. Leiden: E.J. Brill.

Apresjan, Jurij D.; Mel'čuk, Igor A. & Žholkovskij, Aleksandr. 1969. Semantics and Lexicography: Towards a New Type of Unilingual Dictionary. *In:* Ferenc Kiefer (ed.), *Studies in Syntax and Semantics*. Dordrecht: Reidel.

Asher, Nicholas & Pustejovsky, James. 1998. *The Metaphysics of Words in Context*. Tech. rept. Bad Teinach Paper.

Barnett, James; Mani, Inderjeet & Rich, Elaine. 1994. Reversible Machine Translation: what to do when languges don't match up. *In:* Tomek Strzalkowski (ed.), *Reversible Grammar in Natural Language Processing*. Dordrecht: Kluwer Academic Publishers.

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

Béjoint, Henri. 1994. *Tradition and Innovation in Modern English Dictionaries*. Oxford: Clarendon Press.

Benson, Morton. 1990. Collocations and General-Purpose Dictionaries. *International Journal of Lexicography*, vol. 3:23 – 34.

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.

Collins, Allan M. & Loftus, Elizabeth F. 1975. A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, vol. 82:407 – 428.

Collins, Allan M. & Quillian, M. Ross. 1969. Retrieval Time from Semantic Memory. *Journal of Verbal Learning and Verbal Behavior*, vol. 8:240 – 247.

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

Copestake, Ann & Briscoe, Ted. 1992. Lexical Operations in a Unification-Based Framework. *In:* James Pustejovsky & Sabine Bergler (eds.), *Lexical Semantics and Knowledge Representation*. New York: Springer Verlag.

Copestake, Ann; Jones, Bernie & Sanfilippo, Antonio. 1992. *Multilingual Lexical Representation*. ESPRIT BRA-3030 ACQUILEX Working Paper. Pisa.

Davey, B.A. & Priestley, H.A. 1990. *Introduction to Lattices and Order*. Cambridge: Cambridge University Press.

Davidson, Donald. 1974. On the Very Idea of a Conceptual Scheme. *In: Inquiries into Truth and Interpretation*. Oxford: Clarendon Press. 1984.

Feyerabend, Paul K. 1962. Explanation, Reduction, and Empiricism. *In:* Feigl & Maxwell (eds.), *Scientific Explanation: Space and Time*. Minnesota: University of Minnesota Press.

Fodor, Jerry A. 1998. *Concepts: where cognitive science went wrong*. Oxford: Clarendon Press.

Fontenelle, Thierry. 1997. *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen: Niemeyer.

Freese, Ralph; Ježek, Jaroslav & Nation, J.B. 1991. *Free Lattices*. Providence: American Mathematical Society.

Frege, Gottlob. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik*, vol. 100:26 – 50.

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

Gleason, H. A. 1962. The Relation of Lexicon and Grammar. *In:* Fred Householder & Sol Saporta (eds.), *Problems in Lexicography*. Bloomington: Indiana University.

Gleason, H.A. 1955. *An Introduction to Descriptive Linguistics*. New York: Holt, Rinehart and Wilson.

Gleason, H.A. 1965. *Linguistics and English Grammar*. New York: Holt, Rinehart and Winston.

Hanks, Patrick. 2000. Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance. *In: Proceedings of the Ninth Euralex International Congress*.

Hendriks, Petra; Taatgen, Niels & Andringa, Tjeerd. 1997. *Breinmakers & Breinbrekers: inleiding cognitiewetenschap*. Amsterdam: Addison Wesley Longman Nederland.

Hjelmslev, Louis Trolle. 1959. Pour une Sémantique Structurale. *In: Essais Linguistiques*. København: Nordisk Sprog- og Kulturorlag. 1970.

Hutchins, William J. & Somers, Harold L. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Jackendoff, Ray S. 1990. *Semantic Structures*. Cambridge: MIT Press.

Jackson, Howard. 1988. *Words and their Meanings*. London: Longman.

Janssen, Maarten. 2001. Bilingual Dictionaries and Lexical Gaps. *In: Proceedings of COMPLEX 2001*.

Janssen, Maarten & Visser, Albert. to appear. Some Words on *Word* or a full-fledged Aristotelian analysis of the concept *Word*. *La Nuova Critica*.

Janssen, Maarten; Jansen, Frank & Verkuyl, Henk. to appear. The Codification of Usage by Labels. *In:* van Sterkenburg (ed.), *Course Book on Lexicography*.

Johnson-Laird, P.N.; Herrman, D.J. & Chaffin, R. 1984. Only Connections: a critique of semantic networks. *Psychological Bulletin*, vol. 96:292 – 315.

Julien, Marit. 2000. *Syntactic Heads and Word Formation*. Ph.D. thesis, Universitet i Tromsø, Tromsø.

Kamps, Thomas. 1997. *A Constructive Theory for Diagram Design and it Algorithmic Implementation*. Ph.D. thesis, Technische Hochschule Darmstadt, Darmstadt.

Katz, Jerrold. 1972. On the General Character of Semantic Theory. *In:* Eric Margolis & Stephen Laurence (eds.), *Concepts: core readings*. Cambridge: Reidel. 1999.

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

Kuhn, Thomas S. 1970. Reflections on my Critics. *In:* Imre Lakatos & Alan Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Labov, William. 1973. The Boundaries of Words and their Meanings. *In:* C.J. Bailey & R. Shuy (eds.), *New Ways of Analyzing Variation in English*. Washington: Georgetown University Press.

Lakoff, George. 1973. Lexicography and Generative Grammar II: Context and Connotation in the Dictionary. *In:* Raven McDavid & Audrey Duckert (eds.), *Lexicography in English*. New York: New York Academy of Sciences.

Lewis, David. 1972. General Semantics. *In:* Donald Davidson & Gilbert Harman (eds.), *Semantics of Natural Language*. Dordrecht: Reidel.

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

Lyons, John. 1995. *Linguistic Semantics: an Introduction*. Cambridge: Cambridge University Press.

Margolis, Eric & Laurence, Stephen. 1999. *Concepts: core readings*. Cambridge: MIT Press.

Martin, Robert. 1983. *Pour une Logique du Sens*. Paris: Presses Universitaires de France.

Martin, Willy & Gouws, Rufus. 2000. A New Dictionary Model for Closely Related Languages: the Dutch-Afrikaans dictionary project as a case-in-point. *In: Proceedings of the Ninth Euralex International Congress*.

McCawley, James D. 1973. Discussion Paper. *In:* Raven McDavid & Audrey Duckert (eds.), *Lexicography in English*. New York: New York Academy of Sciences.

McNamara, Timothy P. & Miller, Diana L. 1989. Attributes of Theories of Meaning. *Psychological Bulletin*, vol. 106:355 – 376.

Mel'čuk, Igor. 1988. Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria. *Internation Journal of Lexicography*, vol. 1:165 – 188.

Mel'čuk, Igor & Wanner, Leo. 2001. Towards a Lexicographic Approach to Lexical Transfer Machine Translation: illustrated by the German-Russian language pair. *Machine Translation*, vol. 16:21 – 68.

Mel'čuk, Igor A. 1993. *Cours de Morphologie Générale*. Vol. 1: Introduction et Première Partie: le Mot. Montréal: Les Presses de L'Université de Montréal.

Mel'čuk, Igor A. 1995a. The Future of the Lexicon in Linguistic Desctiption and the Explanatory Combinatorial Dictionary. *In:* Ik-Huan Lee (ed.),

Linguistic in the Morning Calm 3: Selected papers from SICOL-1992. Seoul: Hansin Publishing Company.

Mel'čuk, Igor A. 1995b. Phrasemes in Language and Phraseology in Linguistics. *In:* Martin Everaert; Erik-Jan van der Linden; André Schenk & Rob Schreuder (eds.), *Idioms: Structural and Psychological Perspectives*. Hillsdale: Erlbaum.

Mel'čuk, Igor A. 1998. Collocations and Lexical Functions. *In:* A.P. Couwie (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Clarendon Press.

Messelaar, P.A. 1990. *La Confection du Dictionnaire Générale Bilingue*. Leuven: Peeters.

Miller, George A. 1990. Nouns in WordNet: a lexical inheritance system. *International Journal of Lexicography*, vol. 3:245 – 264.

Miller, George A. 1998. Foreword. *In:* Christiane Fellbaum (ed.), *Wordnet: an Electronic Lexical Database*. Cambridge: MIT Press.

Nida, Eugene A. 1958. Analysis of Meaning and Dictionary Making. *Internation Journal of American Linguistics*, vol. 24:279 – 292.

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

Nunberg, G. 1979. The Non-Uniqueness of Semantic Solutions: Polysemy. *Linguistics and Philosophy*, vol. 3:143 – 184.

Ooi, Vincent B. Y. 1998. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Ostler, N. & Atkins, B. T. S. 1992. Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules. *Pages 87–100 of:* J. Pustejovsky & S. Bergler (eds.), *Lexical Semantics and Knowledge Representation: Proc. of the First SIGLEX Workshop*. Berlin, Heidelberg: Springer.

Paasch, Heinrich. 1901. *From keel to truck: marine dictionary in English, French and German*. Antwerpen: Paasch.

Partee, Barbara. 1975. Montague Grammar and Transformational Grammar. *Linguistic Inquiry*, vol. 6:203 – 300.

Pottier, Bernard. 1965. La Définition Sémantique du Lexique Français. *Travaux de Linguistique et de Littérature*, vol. 3:33 – 39.

Pottier, Bernard. 1980. Sémantique et Noémique. *Annuario de Estudios filológicos*, vol. 3:169 – 177.

Priß, Uta E. 1996. *Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases*. Ph.D. thesis, Technische Hochschule Darmstadt.

Pustejovsky, James. 1995a. *The Generative Lexicon*. Cambridge: MIT Press.

Pustejovsky, James. 1995b. Linguistic Constraints on Type Coercion. *In:* Patrick Saint-Dizier & Evelyn Viegas (eds.), *Computational Lexical Semantics*. Cambridge: Cambridge University Press.

Putnam, Hilary. 1981. *Reason, Truth, and History*. Cambridge: Cambridge University Press.

Quillian, M. 1968. Semantic Memory. *In:* Marvin Minsky (ed.), *Semantic Information Processing*. Cambridge: MIT Press.

Quine, Willard Van Orman. 1960. *Word and Object*. Cambridge: MIT Press.

Quine, W.V.O. 1969. Natural Kinds. *In:* Schwartz (ed.), *Naming, Necessity and Natural Kinds*. Ithaca: Cornell University Press. 1977.

Quine, W.V.O. 1981. On the Very Idea of a Third Dogma. *In: Theories and Things*. Cambridge: Harvard University Press.

Rastier, François. 1987. *Sémantique Interprétative*. Paris: Presses Universitaires de France.

Rastier, François; Cavazza, Marc & Abeillé, Anne. 1994. *Sémantique pour l'analyse, de la Linguistique à l'Informatique*. Paris: Masson.

Rey-Debove, Jacqueline. 1971. *Étude Linguistique et Sémiotique des Dictionnaires Français Contemporains*. Den Haag: Mouton.

Rosch, Eleanor & Mervis, Carolyn B. 1973. Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, vol. 7:573 – 605.

Saussure, Ferdinand de. 1972. *Cours de Linguistique Générale*. Paris: Édition Payot.

Scholfield, Phil J. 1979. On a Non-Standard Dictionary Definition Schema. *In:* Hartmann (ed.), *Dictionaries and their Users*. Exeter: University of Exeter. 1979.

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.

Soergel, Dagobert. 1998. WordNet Book Review. *D-lib Magazine*, Oktober.

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

Svénsen, Bo. 1987. *Handbok i Lexicografi*. Stockholm: Norstedts Tryckeri.

Svénsen, Bo. 1993. *Practical Lexicography*. Oxford: Oxford University Press.

Tanguy, Ludovic. 1997. *Traitement Automatique de la Langue Naturelle et Interprétation: Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétative.* Ph.D. thesis, Université de Rennes 1, Rennes.

Verkuyl, Henk J. 1994. *Knowledge Representation in Dictionaries*. Keynote Lecture 6th Euralex International Congress Amsterdam.

Vermazen, Bruce. 1967. Review of Jerrold J. Katz and Paul M. Postal, "An Integrated Theory of Linguistic Description" and Jerrold J. Katz "The Philosophy of Language". *Synthese*, vol. 3:350 – 365.

Visser, Pepijn R.S. & Tamma, Valentina A.M. 1999. An Experience with Ontology-Based Agent Clustering. *In:* Benjamins; Chandrasekaran; Gomez-Perez; Guarino & 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*.

Vossen, Piek & Copestake, Ann. 1993. Untangling Definition Structure into Knowledge Representation. *In:* Ted Briscoe; Valeria de Paiva & Ann Copestake (eds.), *Inheritance, Defaults, and the Lexicon*. Cambridge: Cambridge University Press.

Vossen, Piek; Díez-Orzas, Pedro & Peters, Wim. 1997. The Multilingual Design of the EuroWordNet Database. *In: Proceedings of the IJCAI-97 workshop Multilingual Ontologies for NLP Applications*.

Wardhaugh, Ronald. 1986. *An Introduction to Sociolinguistics*. Oxford: Blackwell.

Weinreich, Max. 1945. YIVO and the Problem of our Time. *Yivo-Bleter*, vol. 25.

Weinreich, Uriel. 1962a. Lexicographic Definition in Descriptive Semantics. *In:* Labov & Weinreich (eds.), *On Semantics*. Philadelphia: University of Pennsylvania Press. 1980.

Weinreich, Uriel. 1962b. Lexicographic Definitions in Descriptive Semantics. *In:* Fred Householder & Sol Saporta (eds.), *Problems in Lexicography*. Bloomington: Indiana University.

Weinreich, Uriel. 1964. Webster's Third: a Critique of its Semantics. *In:* Labov & Weinreich (eds.), *On Semantics*. Philadelphia: University of Pennsylvania Press. 1975.

Whorf, Benjamin L. 1936. The Punctual and Segmentative Aspects of Verbs in Hopi. *In:* J.B. Carroll (ed.), *Language, Thought and Reality: selected writings of Benjamin Lee Whorf*. Cambridge: MIT Press. 1956.

Whorf, Benjamin L. 1956. *Language, Thought, and Reality*. New York: Wiley.

Wierzbicka, Anna. 1972. *Semantic Primitives*. Frankfurt: Athenäum.

Wierzbicka, Anna. 1980. *Lingua Mentalis: The Semantics of Natural Language*. Frankfurt: Athenäum.

Wittgenstein, Ludwig. 1953. *Philosophische Untersuchungen*. Frankfurt am Main: Suhrkamp. 1984.

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

Žholkovskij, Aleksandr & Mel'čuk, Igor. 1965. О Возможном Методе и Инструментах Семантицеского Синтеза. *Науцно-Техницеская Информация*, vol. 5:23 – 28.

Index