An Interlingual Lexical Organisation Based on Acceptions, From the Parax Mock-up to the NADIA System

Gilles Sérasset

ICLA'94, 26-28 July 1994, Penang, Malaysia, pp 20-33.

An Interlingual Lexical Organisation Based on Acceptions

From the PARAX mock-up to the NADIA system

Gilles Sérasset
GETA, IMAG-campus (UJF & CNRS)
BP 53, 38041 Grenoble Cedex 9, France
Gilles.Serasset@imag.fr

Abstract

Many projects are conducted to develop multilingual lexical databases. Some of these projects use an interlingual approach (KBMT-89, EDR, …), where others choose a bilingual approach (Multilex, …).

This paper presents an interlingual approach based on acceptions (word-senses) aiming at the development of a multilingual lexical database management system: NADIA. With this approach, the interlingual lexicon is the union of the acceptions of the languages that appear in the database.

The interlingual set of relations is freely defined by the linguist, with the exception of the predefined "is-a" relation, necessary because acceptions of terms from one language don't always have lexical equivalents in another language.

To illustrate this lexical organisation, we show in detail a mock-up build as a preliminary experimentation: the PARAX mock-up which handles 5 languages: French, English, German, Russian and Chinese.

This lexical organisation has no influence neither on the linguistic content of the monolingual dictionaries, nor on the format adopted by the linguist to encode the information.

The NADIA system is designed to be multilingual (multilingual databases can be handled), application independent (databases can be used for several purposes), generic (the linguistic structures are defined by the linguist) and theory independent (many computational formalisms can be used to define the linguistic structures). The system provides many tools (browser, editor, defaulter, coherence checker, …) to simplify the management of a multilingual lexical database. Moreover, it handles the management of the interlingual dictionary as much as possible.

Finally, we justify our choices by a detailed study of existing projects on multilingual lexical databases and some of the problems they have had to deal with:

EDR and KBMT-89 projects (which use a knowledge based approach) are faced with theoretical and methodological problems (concept discrimination, knowledge representation, concept classification, …). The acception based approach is a good choice to avoid the complexity introduced by knowledge representation problems and to keep the advantages of the interlingual approach.

Multilex provides a way for the linguist to define the different linguistic structures, however, the linguist can only use typed feature structures to code his linguistic theory. We have chosen to provide the linguist with different formalisms (trees, graphs, typed feature structures, sets, …).

Keywords

Multilinguism, Lexical Databases Management System, Lexical Database, Interlingua, Acception.

ICLA'94, 26-28 July 1994, Penang, Malaysia, pp 20-33.

Introduction

The needs in large scale lexical resources for Natural Language Processing (NLP) in general and for Machine Translation (MT) in particular, increases every day. These resources are considered as representing the most expensive part of an NLP system. Hence, we can observe an increasing interest for the development of reusable dictionaries.

To develop a Multilingual Lexical Database, one can use two main approaches. First, the *transfer approach* where the links between the languages are realised via unidirectional bilingual dictionaries. This approach is used by many MT systems and by some lexical database projects (notably Acquilex or Multilex). Second, the *interlingual approach* where the links between the languages are realised via an unique interlingual dictionary. The KBMT-89 project (Knowledge Based Machine Translation) at Carnegie Mellon University in US and the EDR (Electronic Dictionary Research) project in Japan use this approach.

We are interested by the problems posed when constructing and using application and theory independent multilingual lexical databases. In this framework, we chose to develop a Lexical Database Management System, the NADIA system. This system is based on an interlingual approach. We chose *acceptions* as interlingual units. The NADIA system provides many tools for the management of multilingual lexical databases. Moreover, This system gives the linguist a great liberty in the choice of the linguistic structures.

This article presents the NADIA project. After a general study on the interlingual approach, we will present the framework of the system. Then, we will give a detailed view of the *acception based approach*. Next, we will have a closer view on the project with a brief description of Parax, a feasibility study, and the architecture of the system. Finally, we will study some lexical database projects. We will see the choices made for these projects and the problems they have had to deal with. We will finish by a justification of our choices, compared to those projects.

I. Acception based approach

1. Interlingual approach

The interlingual approach uses an artificial intermediary language (called *interlingua* and used as a pivot language) to realise the link between the languages.

An interlingua is an intermediate artificial language in which the utterances of all considered languages can be represented in a neutral way. Hence, an interlingua must have its proper lexicon and its own set of attributes and relations.

An interlingua must be defined with reference to a certain set of natural languages, unless a fixed universe of reference (“ontology”) is represented autonomously in the computer.

An interlingua basically consists in two distinct parts, a lexicon and a set of attributes and relations.

1.1. Interlingual lexicon

The first part of an interlingua is the interlingual lexicon, which must be sufficient to represent the word senses of all considered languages.

Therefore the interlingual lexicon contains at least as many word-senses as each monolingual lexicon.

As an interlingua is defined to establish a link between languages, the interlingual lexicon must provide a lexical link between words of different dictionaries. Hence, two equivalent word-senses in different languages must refer to the same interlingual unit.

However, there is not necessarily a direct mapping between word-senses of different languages. As an example, the French words “rivière” and “fleuve” (in their most common sense) are translated in English by an unique word-sense (the most common sense of the word “river”). The French words are not equivalent[1]. These two word senses are not distinguished in English. A link must be established via the interlingual lexicon between these word senses. However, such a distinction is only pertinent in an English-French context. In an English-Japanese context, this distinction is not relevant, as the Japanese word for river (“kawa”) has the same sense as the English word “river”.

If the interlingua is defined via an autonomous fixed universe or reference (“ontology”), one must give a description of each word-sense of each language in this universe.

In that case, equivalent word-senses of different languages must have exactly the same description. Moreover, “close” utterances (like the French words “rivière” and “fleuve”) must have “close” descriptions. This “distance” between the descriptions must be recognised automatically for translation matter in case of non direct equivalence.

1.2. Interlingual attributes and relations

The second part of the interlingua is its set of attributes and relations. These attributes and relations must be sufficient to code the linguistic particularities of all considered languages.

This part is not easy to define, even if fundamental linguistic studies are slowly producing more and more interlingual or universal “microtheories” (in the words of [Nirenburg and Defrise 1990]) for parts of languages, such as aspect, time, modality, etc., which 20 years ago seemed to be describable only on a language specific manner.

1.3. Existing interlingual projects

Some projects have adopted the interlingual approach in lexical database definitions. Among these projects, we will look at the American KBMT-89 project and the Japanese EDR project. These projects have different aims. The former develops dictionaries for a particular application (a translation system based on knowledge representation) when the later develops lexical databases for any kind of MT system.

KBMT-89 project ([Gates et al. 1989, Nirenburg 1989]) has defined and implemented a translation system based on knowledge. For this purpose, a dictionary of 900 lexical units for English and 800 lexical units for Japanese (representing about 1500 concepts) has been developed.

EDR project ([EDR 1993]) deals with large scale resources. It develops large dictionaries of about 300,000 words in both English and Japanese (200,000 of general vocabulary, 100,000 of terminological vocabulary). EDR also develops a dictionary of 400,000 concepts, a dictionary of 300,000 co-occurrences (both in English and Japanese) and a dictionary of 300,000 bilingual entries (both for Japanese-English and English-Japanese).

---

[1] A “rivière” is a rather small river flowing into another river. A “fleuve” is a large river flowing to the see.

2. Framework of the NADIA project

Our project, NADIA (Neutral Advanced Dictionaries by Interlingual Acceptions), aims at the development of a management system for multilingual lexical databases.

In this framework, we have four main objectives:

- **Multilinguism:** the lexical databases management system must handle multilingual databases. Hence, it must handle various writing systems, various sorting procedures, etc.
- **Application independence**: the management system should not introduce any restriction on the applications that will use the databases. Many systems can use the defined databases (Machine Translation, Natural Language Processing, Natural Language Learning, …) provided that the necessary linguistic information is present.
- **Genericity**: the linguistic structures used by the databases are to be defined by a linguist via a specialised language.
- **Theory independence**: the management system should not introduce restrictions on the linguistic theory used by the databases. On the contrary, it should allow the use of numerous linguistic theories which make use of many computer formalisms.

We have chosen an interlingual approach because we think that it is the best solution to the application independence criterion. Such an approach is compatible with interlingual and transfer approaches as it is possible to generate transfer dictionaries for each language pair present in the database.

In order to reduce problems of interlingual approach, we have chosen to use acceptions as interlingual units. A language acception is a generally accepted meaning of a word. An interlingual acception is one of the acceptions of one of the languages of the database.

The ULTRA project at CRL (Computing Research Laboratory, New Mexico State University) uses an equivalent approach ([Farwell et al. 1992]).

3. NADIA's approach to interlingua

In order to provide useful and powerful tools, the management system introduces some restrictions on the databases themselves: the multilingual lexical databases will be based on acceptions.

3.1. Lexical architecture

The lexical architecture deals with the organisation of the different dictionaries and their inter-relations inside a database.

We stress the fact that this lexical architecture has no influence on the linguistic content of dictionaries, neither on the format adopted by the linguist to code the information.

A lexical database defined within NADIA consists of two kind of dictionaries. The first kind consists of only one dictionary: the interlingual dictionary. The second kind of dictionaries consists of any number of monolingual dictionaries.

The interlingual dictionary contains the acceptions of the different languages of the database.

The monolingual dictionaries contain the linguistic information at the different lexical entries. The linguist is free to define, for each language, his own lexical units and lexical entries, provided that the monolingual dictionary makes the link between entries and acceptions. A monolingual dictionary generally consist of two parts; the first one dealing with acceptions of the language (the purely monolingual part) and the second

one (called contrastive part) dealing with interlingual acceptions that do not have any equivalent in the language of the dictionary.

3.2. Interlingual lexicon

By using acceptions, we avoid the problem of word-sense refinement. For each language, we can choose an existing reference dictionary. Then, the acceptions of the language will be the acceptions found in the reference dictionary.

The interlingual lexicon is the union of all the acceptions of all the languages of the database. If there is a direct correspondence between two acceptions of different languages, these acceptions are merged in an unique interlingual acception. If there is no direct correspondence, the acceptions of a language are kept in the interlingual dictionary.

The management of an interlingual lexicon is a complex task. Hence, this management is done by the machine. The modifications on the interlingual lexicon is done when the machine detects a problem.

3.3. Interlingual set of relations and attributes

The interlingual set of relations and attributes is not fixed. It will be defined by a linguist for each lexical database.

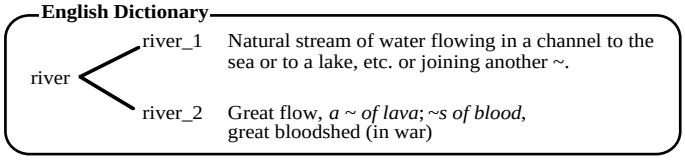
However, in order to handle non direct mappings (see example of “rivière”, “fleuve” and “river” above), a kind of “is-a” relation (relation from super-acception to sub-acception) is defined between acceptions. Then, with this relation, we code the fact that the usual acceptions of “rivière” and “fleuve” are sub-acceptions of the usual acception of “river”.

This kind of relation will always be defined in the lexical databases. It constitutes the minimal set of interlingual relations.

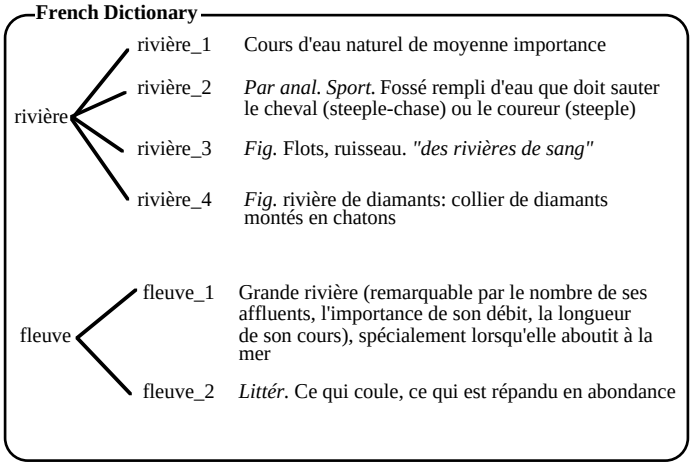
3.4. An example

Let’s consider the above example more in detail.

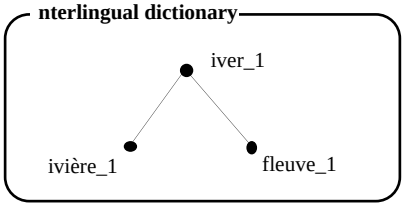
In the purely monolingual part of the English dictionary, the entry “river” has two acceptions:



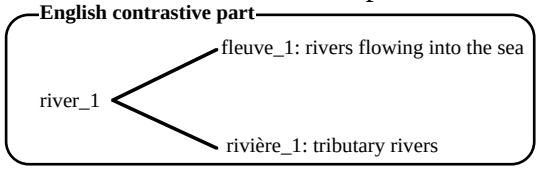
The purely monolingual part of the French dictionary contains the following acception for the entries “rivière” and “fleuve”.



As the English word “river” can be translated by “rivière” or by “fleuve” depending on what it represents, the interlingual dictionary must state that the acceptions “rivière_1” and “fleuve_1” are sub-acceptions of the acception “river_1”.



The contrastive part of the English dictionary reflects this problem and contains the information allowing the distinction of the two acceptions.



II. The NADIA project

1. Parax

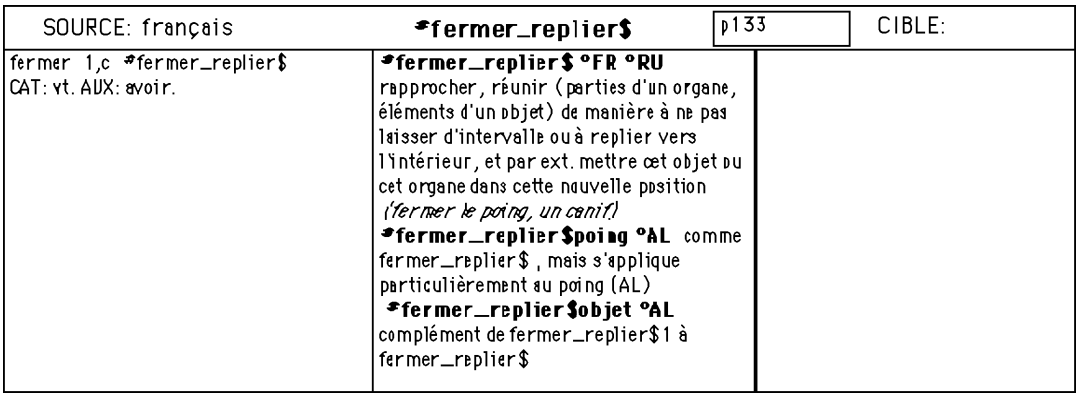
Parax is a mock-up implemented with Hypercard™ by Étienne Blanc to study the feasibility of the acception based approach. This mock-up currently handles 5 languages: French, English, German, Russian and Chinese.

The linguistic structure of monolingual dictionaries comes from the dictionaries of the ARIANE system: a translator generator developed at GETA [Boitet et al. 1982, Boitet et al. 1985] . This linguistic structure is fixed.

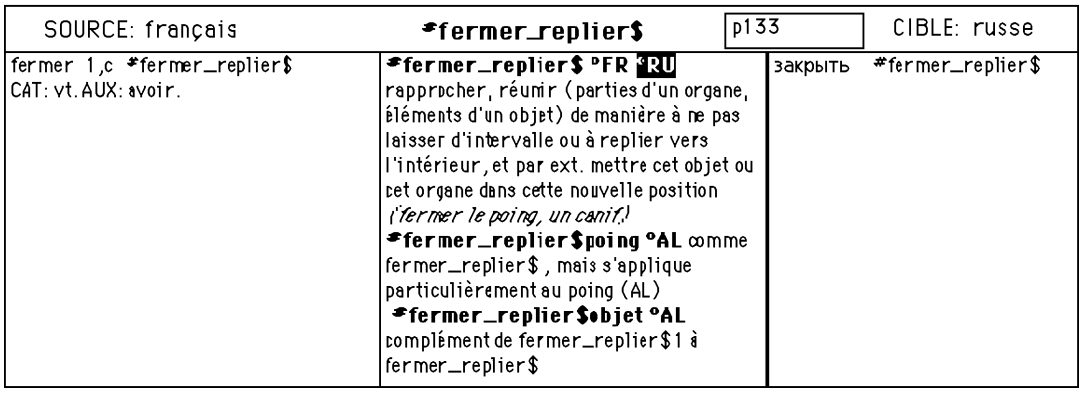
The mock-up illustrate the choice of an interlingua by acceptions. The user can select a word in one of the languages. Then, Parax displays a list of acceptions corresponding to the chosen word. Each acception comes with a definition in French (the working language of the database).

	fermer_1
fermer_1,a #fermer_fermeture$ CAT: vt. AUX: avoir.	**fermer_fermeture$ °AL °AN °FR °RU** appliquer une partie mobile pour boucher un passage, une ouverture *(" fermer la porte, les rideaux.)* **fermer_fermeture$organe °CH** fermer *("bouche, yeux,...,)* **fermer_fermeture$norg °CH** fermer *("sauf bouche, yeux...,)*
fermer_1,b #fermer_lieu CAT: vt. AUX: avoir.	**fermer_lieu °AL °AN °FR °RU** priver de communication avec l'extérieur, par la mise en place d'un élément mobile; interdire le passage *("fermer sa chambre, une valise, une route.)*
fermer_1,c #fermer_replier$ CAT: vt. AUX: avoir.	**fermer_replier$ °FR °RU** rapprocher, réunir (parties d'un organe, éléments d'un objet) de manière à ne pas laisser d'intervalle ou à replier vers l'intérieur, et par ext. mettre cet objet ou cet organe dans cette nouvelle position *("fermer le poing, un canif.)* **fermer_replier$poing °AL** comme fermer_replier$, mais s'applique particulièrement au poing (AL) **fermer_replier$objet °AL** complément de fermer_replier$1 à fermer_replier$

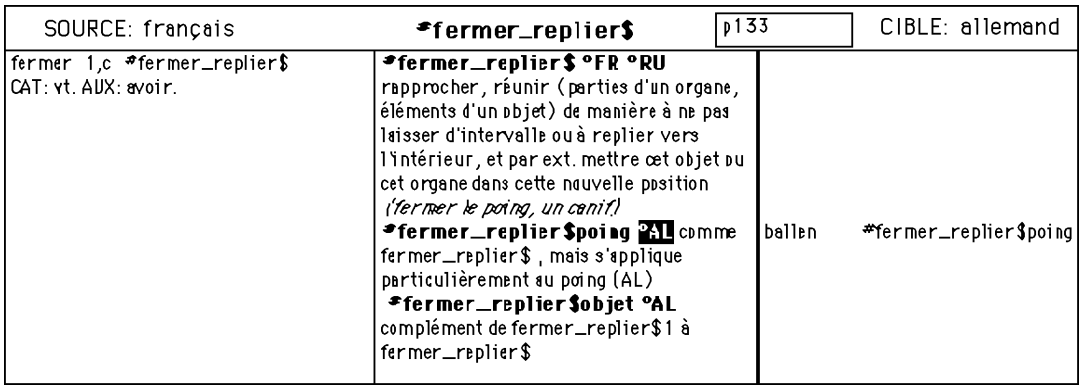
Then, the user selects one acception. In the given example, there is a contrastive lexical problem between French and German. Two German acceptions correspond to the selected French acception: “fermer_replier” (closing by folding something). If the object of the verb is a fist, the German translation is “ballen”, else, it is “zuklappen”.



If the user wants a Russian equivalent, there is a direct equivalent to the selected acception.



However, if he wants the German translation, he has to chose the desired sub-acception.



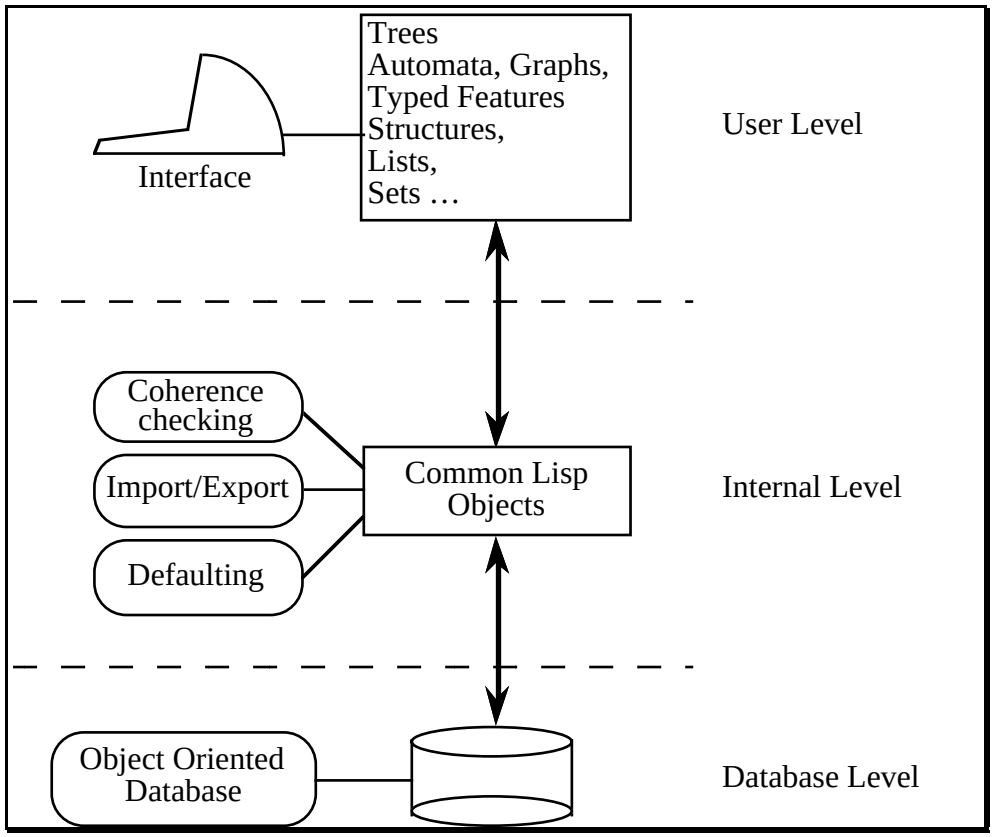
2. NADIA : software architecture

The Parax mock-up is an illustration of the interlingual approach we chose. This mock-up does not satisfy the theory independence objective. It does not provide specialised tools for the management of multilingual lexical databases.

In the framework of the NADIA project, we will develop a prototype of such a system using Common Lisp and Common Lisp Object System (CLOS).

NADIA’s software architecture contains 3 parts:

- *The database level* is the lowest one. It is the level where objects are stored/restored. This level will remain completely transparent for users (lexicographers, linguists, end-users…).
- *The internal level* is the level to which the different tools have access. The different linguistic objects are coded as Common Lisp Objects. This level will remain transparent for users.
- *The user level* is the abstract level at which the user works. At this level, the user directly manipulates linguistic object (trees, typed feature structures, lists, automata, graphs, sets,…). The tools will provide outputs and accept inputs at this level.



The functioning is based on the backward and forward motion between the user, the internal and the database levels.

A linguist defines the structures of the entries at the user level using usual (trees, features structures, graph, …). This definition is translated in terms of Common Lisp Objects.

Definitions and integrity rules are defined by a linguist at the user level. These rules are "compiled" and effectively used at the internal level and permanently stored in the database.

Queries are defined at the user level. They can be translated and evaluated at the internal level. The results of the queries will be presented at the user level.

The user can import or export entries or part of entries. The importing and exporting definitions will be defined at the user level, "compiled" and effectively used at the internal level. The result of exporting procedures will be stored in a SGML (Standard Generalised Markup Language) format reflecting the linguistic structures (TEI (Text Encoding Initiative) guidelines will be followed as much as possible).

3. Management of the dictionaries

There are 2 kinds of dictionaries in a NADIA multilingual lexical database:

- *The monolingual dictionaries* are divided in two parts:
  - a purely monolingual part that contains acceptions of the language and the linguistic information that concern them
  - a contrastive part that contains information about acceptions existing in an other language, but not in the language of the dictionary.
- *The interlingual dictionary* contains the interlingual units (acceptions) and their relations.

The linguist defines the linguistic structures of the entries for each monolingual dictionary via a specialised language. It is possible to define coherence, integrity and defaulting rules on these structures. The linguist manages the information of monolingual dictionaries with tools provided by the NADIA system (editors, defaulters, coherence checker, …).

The interlingual dictionary is complex to manage. An interlingual acception is created when a new acception appears in one language. The interlingual lexicon must provide lexical links between the different languages. For that purpose, all equivalent acceptions must be linked to the same interlingual acception. Hence, when adding a new term, the lexicographer has to check if an equivalent interlingual acception already exists in the interlingual lexicon. If so, he has to link the term to the acception, if not, he has to create a new acception in the interlingual lexicon. Such a basic management is not at all easy as it implies some constraints:

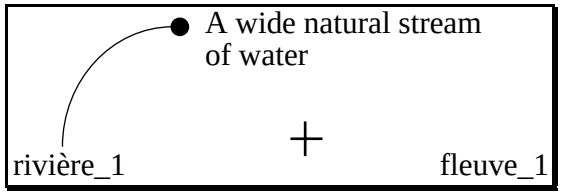
- The lexicographer has to browse the interlingual database to determine if the acception has been already defined by another lexicographer. This detection is not easy as the database may not be fully complete.
- All problems detected in the interlingual dictionary have to be solved in a coherent way.
- The interlingual lexicon must provide information for the lexicographer who checks the existence of an acception. This information must be sufficient to define unambiguously an acception and must be understable for all lexicographers. So, it must be given in a common language.
- The lexicographer who has to create a new acception must provide this information, using the common language (which is not necessarily his native language).

For these reasons, the management of the interlingual dictionary will be handled by the system as much as possible. The system will create a new acception in the interlingual lexicon when a new acception is created in a monolingual dictionary. The system will link a new term with an already existing acception when it is possible.

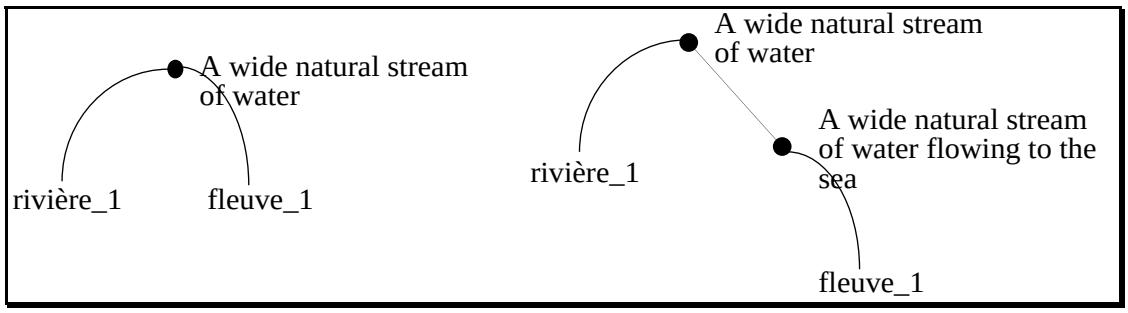
Determining the existence of an interlingual acception or the need to create a new one is done by a “side effect”. The lexicographer gives an equivalent of a term. In this way, the system knows if the corresponding interlingual acception exists or not and creates a new acception if necessary.

For example, let’ us have a look at the indexation of the French words “rivière” and “fleuve”. We suppose that the lexical database already contains an English dictionary.

The lexicographer indexes the French word “rivière”. He finds the acception corresponding to the word “river” in English and links the new term to this already existing acception. After a while, he has to index the word “fleuve” and he finds the acception corresponding to “river”.

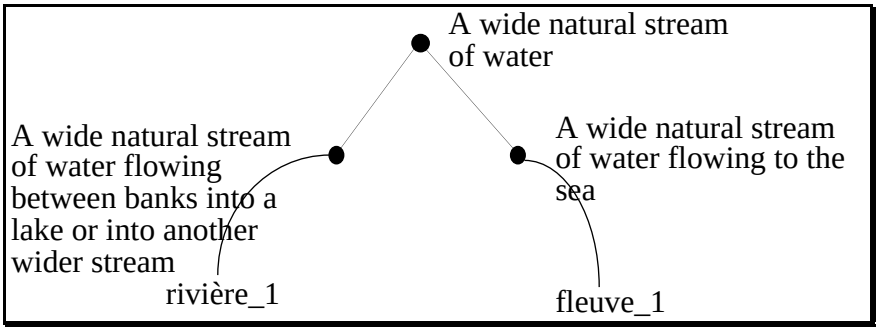


In front of this problem, the lexicographer might adopt different solutions: link this acception with the entry “fleuve” or create a new acception for “fleuve” and link it with an “is-a” relation to the acception of “river”.



The first choice is incorrect as the acception is already linked to “rivière” and as “rivière” and “fleuve” are no synonyms. The second choice is also incorrect as a “fleuve” is not a “river”.

The correct choice to resolve this problem is to modify the link between “rivière” and the original acception (which corresponds to “river”) with the creation of two new acceptions (one for “rivière”, one for “fleuve”). These new acceptions are sub-acceptions of the original one.



In order to ensure a coherence in the management of the interlingual dictionary, and to simplify the work of the lexicographers, it is the system that takes in charge the different problems that can arise. It will detect them and propose solutions to the lexicographer.

III. Justification of the choices

1. Analysis of existing projects

1.1. KBMT-89

KBMT-89 (Knowledge Based Machine Translation) is a research project of Carnegie Mellon University’s Center for Machine Translation. The goal of this project is to construct a MT system based on knowledge. The KBMT-89 system uses a knowledge representation of the domain as an interlingua.

The knowledge acquired by the system includes an ontology (about 1500 concepts). The lexicons contain about 800 Japanese lexical units and 900 English lexical units. A more detailed description of the project can be found in [Meyer et al. 1990, Nirenburg 1989].

KBMT-89's dictionaries are composed of:

- a knowledge base (ontology),
- monolingual lexicons (Japanese and English).

The knowledge base contains a set of concepts, classified in a hierarchy. These concepts are used as an interlingua.

The lexicons contain a set of "super-entries" (lemmas) decomposed in entries (senses). Each of these senses are combined with linguistic information (category, orthography, phonology, morphology, syntactic features, syntactic structures, semantics, lexical relations and pragmatics).

Each sense of a word is generally linked to a concept of the knowledge base. A sense can be linked to an interlingual structure which is not reflected by a concept of an ontology (an attitude or a relation).

1.2. EDR

EDR (Electronic Dictionary Research) is a Japanese project aiming at the construction of a large scale lexical database for different MT systems. EDR has developed word dictionaries, a concept dictionary, bilingual dictionaries and co-occurrence dictionaries for English and Japanese.

The EDR concept dictionary contains a description of the concepts. This description is done by way of relations between concepts. EDR uses 32 types of relations and 5 attributes for this description. The concepts are also classified in a hierarchy.

The concept dictionary is used as an interlingual dictionary. Word dictionaries are linking entries and concepts and contain the linguistic information attached to the entries.

1.3. Multilex

The European project Multilex (DG XIII - ESPRIT project) aims at the definition of standards for the construction of multilingual lexical databases. Multilex deals only with languages of the European Community.

Multilex uses a transfer approach. A Multilex lexical database contains 2 kind of dictionaries: monolingual dictionaries and bilingual dictionaries.

The monolingual dictionaries contain lexical units (senses) and their associated linguistic information. A language has been developed for the linguist to define the structures used in a dictionary. These structures must be coded as Typed Feature Structures.

Bilingual dictionaries are unidirectional. Each bilingual unit is composed of a source lexical unit, a target lexical unit, a condition that is required to select this bilingual unit and a transformation rule linking the structures of the source lexical unit and the structure of the target lexical unit.

2. Encountered problems

It is interesting to analyse the main problems encountered by these projects. All of these projects face problems with the coding of linguistic information and the interlingua based project faces problems with concept refinement. We will also see the problems induced by knowledge representation and problems induced by concept classification.

2.1. Representation of linguistic information

The EDR project represents linguistic information as attributes. The possible attributes are defined a priori and cannot be modified. The linguistic information is not very detailed, but the present information is sufficient to be used by MT systems.

The Multilex project is more flexible as it allows the linguist to define the linguistic structures of the dictionaries. However, the linguist is obliged to use Typed Feature Structures to code his lexical information. These structures are well adapted for some problems but they are not well adapted to code certain linguistic structures like trees or automata.

2.2. Concept refinement

In order to use concepts as an interlingua, one must answer the question: “what is a concept?”. Let’s take an example:

- An elephant appears.
- An elephant is a clever animal.
- The elephant is an endangered species

One has to know if there are 3 different concepts for “elephant” or if there are 3 different realisation of the concept “elephant”.

EDR considers that the 3 sentences contain 3 different realisations of the same concept (the first sentence is about an individual elephant, the second one is about a typical elephant and the third one is about the species).

KBMT-89 has defined four criteria to refine the concepts. One of these criteria states that if a lexical unit has two incompatible sets of grammatical attributes, then two lexical units must be created. The grammatical attributes can be morphological, syntactic or lexical. But this criterion is not absolute: it depends on the size and systematicity of the differences between the 2 sets.

In the given example, it is not easy to see if we deal with one or three concepts. In the first sentence, the noun “elephant” is a discrete noun (it is possible to say “3 elephants appear”). In the second sentence and the third sentence, “elephant” is not discrete (saying “3 elephants are clever animals” is not analogous). Moreover, in the third phrase, one can not use the article “an”.

One must make a choice between creating 3 concepts or only one. When this choice is done (whatever it is), the same choice must be done when an analogous problem appears. When constructing a real scale database, many lexicographers will have to work together. Ensuring the coherence of the choices is a complex methodological problem.

2.3. Problems introduced by the knowledge representation

EDR and KBMT-89 use a knowledge representation as an interlingua.

This choice introduces some methodological problems when one wants to construct real scale lexical databases.

First, it is difficult to describe concepts for a large scale (EDR, which successfully developed 400,000 concepts, needed 1200 man-years). To be really feasible, such a description should be done via a (semi-)automatic extraction of concepts. EDR has largely used this approach (from a large corpus).

When extracting knowledge, EDR has used a large corpus of texts. For speed and complexity reasons, EDR had to make hasty generalisations [H. Suzuki, personal communication]. Let’s take for example the sentence “An elephant eats an apple”. This

sentence states that a particular elephant is eating a particular apple. EDR extracts two relations from this sentence:

- Any elephant can be an agent of the verb “to eat”,
- Any apple can be the object of the action “to eat”.

2.4. Problems introduced by a concept classification

EDR and KBMT-89 reduce the number of relations between concepts by a factorisation. This factorisation is done by way of a hierarchy and an inheritance mechanism.

For example, in the EDR hierarchy, the fact that a bird can fly is coded. The inheritance mechanism states the fact that a lark, a swallow, … can fly.

However, this classification of concepts can introduce theoretical and methodological problems.

First, one has to deal with exceptions. How can we represent the fact that an ostrich is a bird that cannot fly? EDR has chosen to declare texplicitly hat an ostrich can not fly via a negative relation [H. Suzuki, personal communication]. This assertion replaces the inherited assertion. However, negative relations should be used carefully and cannot be inserted automatically as they can introduce incoherence in the database.

The modification of such a hierarchy is really difficult as all the sub-concepts of the modified concept will be affected.

The addition of a relation to a concept in the hierarchy can be problematic if this concept has sub-concepts. This new relation will be inherited by all the sub-concepts. A verification will have to be done in order to verify if there are no new exceptions among the sub-concepts.

3. Justification of the choices

The choices we have made for the general organisation of the project are justified by our desire to ensure a certain compatibility (at least for data exchange) with the above projects and by the problems they encountered.

Multilex uses a transfer approach, when EDR and KBMT-89 have chosen an interlingual approach. In order to be compatible with both approaches, we have chosen an interlingual approach. With such an approach, it is possible to generate transfer dictionaries (compatible with Multilex dictionaries).

The EDR project uses a relatively fixed linguistic structure. Multilex is more adaptable, but allows only the use of Typed Features Structures. As adaptability is a really strong point, we chose to give the linguist the possibility to define his own linguistic structures. To increase this adaptability, we do not restrict the choice of the computer structures that will code the linguistic structures. The NADIA system allows the use of most of the linguistic structures used now in Natural Language Processing. It is also possible to mix these different structures in a single entry, in order to dispose of an representation adapted to each linguistic problem. This way, we guarantee an independence of the system toward the chosen linguistic theory.

The two studied interlingual projects use a knowledge representation approach. We saw that this approach introduces some theoretical and methodological problems. In order to avoid some of those problems, we chose to use the acceptions as interlingua. This choice avoids the problem of refinement of the interlingual units (this refinement is systematic as we can use existing dictionaries as a reference). We also do not have to represent knowledge which lowers the cost of the creation of such a base.

The EDR project uses a classification of concepts to reduce the number of relations. As we do not use a description of concepts, this classification is useless. We use only one

relation between acceptions: a relation from super-acception to sub-acception. This relation is useful to code contrastive problems. It is not intended to code a complete hierarchy. It will only be used “locally” in case of contrastive problems.

Conclusion

We have presented here the main ideas of the NADIA project, which aims at the construction of a management system for multilingual lexical databases.

The NADIA project uses an original interlingual approach: interlingua by acceptions. This approach avoids some problems of knowledge representation that often appear when dealing with interlingual projects.

By giving the linguist the possibility to choose the linguistic structure(s) he wants, we guarantee an independence toward the linguistic theory used for each base.

The NADIA prototype described here is currently under development. This prototype will open the way to the solution of different problems. For example, improving import/export tools will bring us to the problem induced by the sharing of linguistic data (especially data implementing different linguistic theories).

We would also like to develop the analogy between dictionaries and structured documents. Such an analogy gives us the ability to use results coming from this field. Hence, we can define different views of a lexicon. It is interesting to study how to simulate different linguistic theories using different views of the same lexicon.

References

Boitet C., Guillaume P. and Quezel-Ambrunaz M. (1982). *ARIANE-78: an integrated environment for automatic translation and human revision*, COLING-82, July 1982: pp. 19-27.

Boitet C., Guillaume P. and Quezel-Ambrunaz M. (1985). *A case study in software evolution: from ARIANE-78.4 to ARIANE-85*, Theoretical and Methodological Issues in Machine Translation of Natural Languages, August 14-16, 1985: pp. 27-58.

EDR (1993). *EDR Electronic Dictionary Technical Guide*, Japan Electronic Dictionary Research Institute Ltd., Project report n° TR-042, August 16, 1993: 144 p.

Farwell D., Guthrie L. and Wilks Y. (1992). *The Automatic Creation of Lexical Entries for a Multilingual MT system*, COLING-92, July 23-28, 1992, vol. 2/4: pp. 532-538.

Gates D. et al. (1989). *Lexicons,* Machine Translation, vol. 4(1): pp. 67-112.

Meyer I., Onyshkevych B. and Carlson L. (1990). *Lexicographic Principles and Design for Knowledge-Based Machine Translation*, Carnegie Mellon University, Technical Report n° CMU-CMT-90-118, August 13, 1990: 66 p.

Nirenburg S. (1989). *Knowledge-based machine translation,* Machine Translation, vol. 4(1) : pp. 5-24.

Nirenburg S. and Defrise C. (1990). *Lexical and Conceptual Structure for Knowledge-Based Machine Translation*, ROCLING III, August 20-22, 1990: pp. 105-130.

-o-o-o-o-o-o-o-o-o-o-