Natural Language Engineering **1** (1): 000–000. Printed in the United Kingdom

MLDS: A Translator-oriented MultiLingual Dictionary System

E. AGIRRE, X. ARREGI, X. ARTOLA,
A. DIAZ DE ILARRAZA, K. SARASOLA
and A. SOROA
Informatika Fakultatea, Computer Science Faculty
P.K. 649, 20080 DONOSTIA, Basque Country, Spain.
e-mail: jiparipx@si.ehu.es

(*Received 16 July 1999*)

Abstract

This paper focuses on the design methodology of the MultiLingual Dictionary-System (MLDS), which is a human-oriented tool for assisting in the task of translating lexical units, oriented to translators and conceived from studies carried out with translators. We describe the model adopted for the representation of multilingual dictionary-knowledge. Such a model allows an enriched exploitation of the lexical-semantic relations extracted from dictionaries. As well, MLDS is supplied with knowledge about the use of the dictionaries in the process of lexical translation, which was elicitated by means of empirical methods and specified in a formal language. The dictionary-knowledge along with the task-oriented knowledge are used to offer the translator active, anticipative and intelligent assistance.

1 Introduction: Machine-Assisted Translation and Dictionaries

Translation support tools have become a novel and promising branch in the field of machine-assisted translation in the last few years. Instead of fully-automatic machine translation, a partially automated way of translating has been promoted in the so-called "workstation approach". Thus, machines would cooperate with humans and they would complement the translators' activity with the purpose of making the translation process easier and the resulting translation better.

In the early 80's Martin Kay advocated a modest view in which machines were gradually allowed to take over certain functions in the overall translation process. He had the view of "Little steps for little feet!"

Along with this view, it has been observed that tools for translation cannot be made without the co-operation of human translators. In fact, machine translation research has not paid enough attention to human translators' behaviour (Johnson and Whitelock 1987; Steiner 1993; Hauenschild 1993). We think that this is a serious methodological fault, and the obvious consequence is that systems are not constructed to accord with the translators' activities. In turn, these tools lead translators to change their behaviour, perhaps undesirably, rather than just assist them, which of itself can create further interaction problems as well. It is therefore necessary to formulate a theory of machine-assisted translation based on the experience of human translation and adapted to it. Such an approach would allow both designers and engineers to better customize the device used for machine-assisted translation. Nowadays commercial tools involved in translation, such as multilingual word processing (Cole et al. 1997), on-line dictionaries (see, for instance, `http://members.tripod.com/~marquardt2/index-2.html`) or translation memories (Dagan and Church 1994), are in widespread use. That is because machines are progressively being adapted to users' needs.

1.1 Our proposal

Our approach, following P. Isabelle (1992), is to start from the study of the human translation process and to design a system based on translator's activities. This principle has been applied in a strict way and it has entirely conditioned our design methodology and, as a consequence, the resulting system.

We are interested in the design of a help system specialised in the specific task of the translation process. Our primary aim is to limit the scope of the system to lexical translation. This idea is motivated by two principles that we noticed when working with human translators: i) the relevance of lexical translation, and ii) the lack of automatic assistance.

From our point of view, lexical choice is an important point in the translation process. Translators are often involved in the task of choosing suitable target lexical units that correspond with those that are in the source text. Such a task may not be easy or even simple and the behaviour of human translators when choosing lexical translations reflects complex cognitive processes. Taking this into account, we consider that a way of assisting translators in this activity is by offering usable lexical knowledge as well as information about the lexical translation task itself.

1.2 The Role of Dictionaries

Dictionaries, which are the most used tool in this process, do not seem able to deal with the problem of the lexical selection in its whole dimension. Neither conventional nor electronic dictionaries treat lexical-level translation as a difficult problem in which rich cognitive capabilities are necessary. It is true that in the electronic age the notion of a dictionary is changing, given that new ways of storing, retrieving, and updating lexical information have been developed. Nevertheless, they hardly conform to the cognitive processes involved in human lexical selection. Hence, we agree with (Fillmore and Atkins 1994) that the traditional lexicographic approach to word meaning must change in response to the demands of translators. A way of adapting could be to incorporate into a dictionary system both lexical knowledge and knowledge about the use of dictionaries when dealing with a lexical problem. In this way, the dictionary system becomes an active tool able to better support the task of lexical translation, rather than being only a repository, more or less structured, of words and definitions.

In our opinion, the idea of the "active" dictionary introduced by (Martin and Al 1990) is particularly relevant : "...the use of a dictionary can be seen as a typical problem-solving activity, and user-orientation should involve both (static) knowledge and dynamic features (strategies, aims, needs) of the intended user". Even considering dictionaries as human user-oriented tools (traditional concept of the dictionary), they could incorporate "dynamic features" by means of appropiate computational functions.

This notion of a dictionary as a dynamic tool is adequate to be applied in the context of machine-assisted translation. Such a dictionary should be useful either when understanding source lexical items, when searching for equivalents, or even when predicting target lexical forms. For that, it needs to incorporate reasoning mechanisms in order to exploit the explicit and implicit information from the dictionary. Each information unit would be considered as a basic object in the global reasoning process. Furthermore, along with the usual information about the meaning of the entries, dictionaries should show how to use words in context. In other words, we advocate that dictionaries should actively cooperate in finding the correct translation.

It could seem that lexical-level assistance is a somewhat limited approach to a translator's workbench. But, on the contrary, we note that the lexical components of natural language processing applications are more and more important nowadays. For that reason, lexical assistance could be an important tool in machine-assisted translation.

The MultiLingual Dictionary system (MLDS) is proposed in this paper as a first phase of implementing our vision of a translation dictionary. In section 2 we describe the general principles that have guided the design of MLDS and its more relevant features. Section 3 discusses the modelling process and describes a model of the use of dictionaries during the translation task. This model is exploited in the MLDS system.

Section 4, describes the knowledge representation model adopted in MLDS to represent the lexical knowledge acquired from source dictionaries. Sections 5 and 6 present a functional view of MLDS. We first introduce the general features of the functionality of the system and then illustrate this functionality with examples. In section 7 some conclusions are drawn.

2 The Dictionary as an Expert System

In machine-assisted translation research, the design is, generally, characterised by what we think computer programs can do, rather than what really goes on when translators translate. We propose that the starting point in the design process should be observation, knowledge elicitation, and modelling of expertise. The resulting model will be transformed and made explicit into a system in the way that modern Knowledge Engineering advocates and supports.

In our application, the design of the active translation-oriented dictionary system involves the use of Artificial Intelligence techniques, given that the desirable active behaviour has to be related to reasoning and deduction processes. From this point of view, we see the dictionary system as an expert system in the lexical translation process.

The starting point of MLDS is the Intelligent Dictionary Help System (IDHS) (Agirre et al. 1994c), which is conceived as a monolingual (explanatory) dictionary system for human use (Artola and Evrard 1992). The system provides various access possibilities to data, allowing the user to deduce implicit knowledge from the explicit dictionary information. IDHS deals with reasoning mechanisms analogous to those used by humans when they consult a dictionary.

The design of MLDS has been structured in several steps as recommended by KADS[1] ("Knowledge Acquisition and Documentation Structuring") methodology (Schreiber et al. 1993). Firstly, we have defined a conceptual model on how translators use dictionaries in order to find correct translations. The conceptual model is the starting point of the design process.

Secondly, the formal specification of the functional behaviour is defined. Finally, the architecture of the system is built in a way that makes the conceptual model executable. This model represents the process involved in the translation of words in context using lexical information.

The KADS design principles assisted in clearly distinguishing the levels of knowledge that MLDS manages. The levels, as defined using KADS, are the following:

- Domain layer. In our system the domain layer has been created from the dictionary contents. Entities and relations extracted from dictionaries have been described in a Dictionary Knowledge Base (DKB), according to relational lexical-semantic principles (see section 4).
- Inference layer. If we consider dictionaries as active tools, it is a logical conclusion to include the basic lexical-semantic inferences within the dictionary knowledge. Thus, dictionary-inferences are usable in different contexts. The domain and the inference layer are organised into different modules of the Dictionary Knowledge Base.
- Task layer. We specify how to combine inferences in order to achieve a given goal. The top-task of our model is the Translation of the Source Word. Tasks are laid out in a hierarchy, and hierarchical relations mean that, in order to achieve goals of the parent-tasks, it is necessary to combine and execute sub-tasks. The task layer reflects the knowledge we have elicited from human translators referring to the way words[2] of the source language are translated into the target language using dictionaries.
- Strategic layer. It describes the way to interpret and dynamically combine the knowledge of the task layer in order to solve the lexical translation problem.

[1] This methodology has been succesfully applied in the design of knowledge-based systems. But, as far as we know, it has not been used in the design of machine-assisted translation system

[2] The term *word* must be understood *lato sensu*. It stands for simple words as well as for multi-word terms or idiomatic expressions

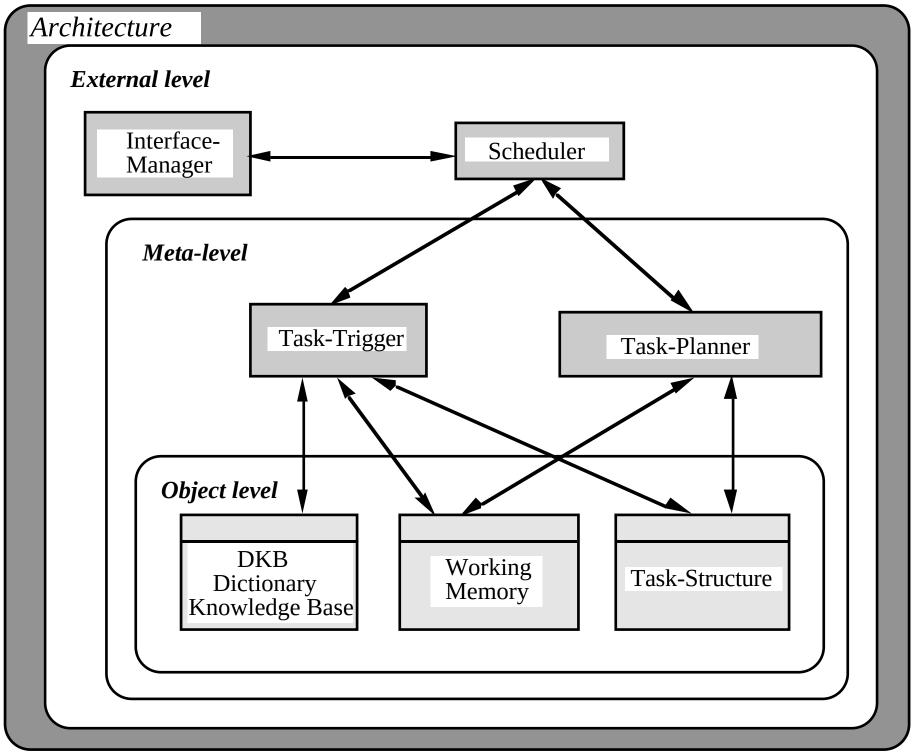

Fig. 1. View of the Architecture of the MLDS system.

The strategic layer is of dynamic nature in the sense that the actual state of the translation process is taken into account when deciding the most suitable task to be activated. Thus, the strategic layer can be seen as the manager and planner of the static knowledge resident in the task layer.

The activities being modelled by these layers should be supported by a specific and modularized architecture. In figure 1, the physical architecture of MLDS is shown. It is composed of seven modules: three of them are data-modules, and the remainder are operational ones.

In the object level we can distinguish three modules:

- The Dictionary-Knowledge Base (DKB), that is, the repository of lexical and semantic knowledge. The domain and inference layers are set in the DKB.
- The task-structure, that contains knowledge about the lexical translation process, corresponding to the task layer. This information is obtained by a modelling process, as described in section 3.
- The working-memory is the storage of the state of the activity. It is used when choosing the task to be executed during the translation process.

Two operational modules interact directly with the object level. They constitute the meta-level of the system:

- The task-planner decides which task must be executed at each moment.

- The task-trigger activates the primitive tasks and updates the working-memory.

In the external level other two operational modules are placed:

- The scheduler manages control and data-flow.
- The interface-manager controls the input/output and the dialogues with the translator.

In section 5, we will explain how these modules work together.

3 Modelling the Use of Dictionaries in Translation

From our view, as tools for translation cannot be satisfactorily designed without the co-operation of human translators, any attempt to incorporate task-dependent behaviour into a dictionary system should begin with a study of the tasks involved and the users' interaction with a dictionary.

Certainly, it would be desirable to have a well-founded theory about this kind of use and interaction. The use of dictionaries has been previously researched from different perspectives, (Hartman 1985; Atkins and Knowles 1990; Nuccorini 1994), however, a general theory has not been presented to date.

Our work has been fundamentally empirical. We have not limited our work to a questionnaire-based method to collect information. With questionnaires, as Hatherall remarks (84), it is not easy to know what subjects are really saying what they do, what they think they do, or what they think they ought to do. As an alternative, we have used both direct observation and personal interviews as presented below:

a) Direct observation protocols. The translators were given several texts to be translated (in our case French and Basque texts) along with several dictionaries (monolingual and bilingual with different characteristics) in order to record their problems, the solutions they adopted, and the tasks they carried out. The aim was to characterise the activity of human translators by observing the task of translating words, expressions, context-dependent phrases and even paragraphs (rarely). Each time the human translator used a dictionary, the unit to be translated, the dictionary used, the consulted dictionary entry and the type of consultation were recorded.

b) Personal interviews with professional translators. These interviews have allowed us to detect different uses of the dictionary according to their experience in the subject. Additional questions were posed to the experts: what characteristics a dictionary should have in order to be useful when translating, their interest about having computerised dictionaries and their expected functionality, and so on.

The data extracted leads us to interesting considerations, although some of them are already well-known: i) both occasional and expert translators need different and adaptable help services, ii) some translators, mainly occasional ones, find bilingual dictionaries very useful, iii) access to multi-word terms is usually unavailable in dictionaries, iv) context is important when translating a text, v) dictionaries for translation should give grammatical and usage information, vi) the proximity between languages is helpful, but attention must be paid to "false friends", and dictionaries should prevent them. These considerations have been implicitly incorporated into our model.

3.1 Other models

One of the conclusions we extracted from our study lies in the fact that the use of the dictionary must be seen as a process, rather than as a specific action. In (Neubach and Cohen 1988) this idea is consistently encouraged, suggesting no complete approach to the use of dictionaries can be made without analysing this process.

Some attempts to model the use of dictionaries in translation have been already carried out. In a research study made in the *School of Translation and Interpreting of Maastricht* (Starren and Thelen 1990), the use of dictionaries is organised in four steps: discovering meaning, finding receptor language equivalents, checking meaning of receptor language item, and formulation of final translation.

Different models of the translation process are presented in (Sager 1994). These models exhibit certain limitations which make them unsuitable for our specific purpose such as:

- they are made in a statically descriptive way which idealises the process. No indication is given of the nature or complexity of the mental processes involved.
- they are theoretically-speculative and not based on empirical data.
- they do not account for task specifications.

In order to manage our approach, we need a model formal enough to be computerised, and furthermore, we want to focus on the translation of word-forms. For these reasons, a different way of modelling lexical translation is proposed.

3.2 Description of the model: a general view

Starting from what we have observed and recorded, the model of the use of dictionaries when translating words has been formalised into three steps i.e., the conceptualisation, the specification and the operationalisation.

In the conceptualisation step the space of the problem was bounded by distinguishing and classifying the entities involved in it. We have classified the entities into three types: objects/roles, subtasks and states.

The specification has consisted of clearly describing the entities and their relations. It is a task-oriented specification, given that tasks can be seen as the core entities where objects and state are described and related.

Finally, the operationalisation has dealt with the way of carrying out the specified tasks. In this step we are concerned with the strategy of the translator, and therefore we have used an algorithmic language.

The model of expertise obtained in the elicitation process is stored in the task-structure module of the system. The knowledge contained in the task-structure

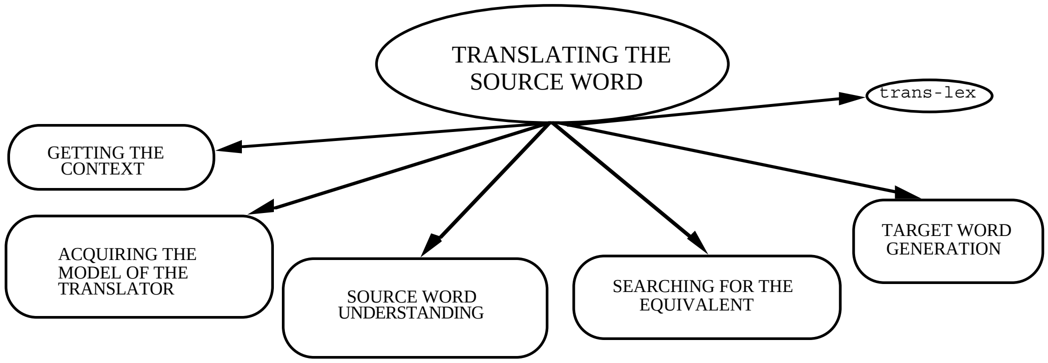

Fig. 2. Top level of the decomposition diagram of the tasks involved in the lexical translation process.

reflects different strategies, tasks and ways of using dictionaries by translators (see Appendix 1, figures 8, 9, 10). Figure 2 shows the top level of such a structure.

In our model, we consider tasks as divided into composite (expressed by uppercase in figures) and primitive ones (lowercase in figures).

The composite tasks are described as non-trivial processes to be decomposed into subtasks when they are carried out. At the same time, each one of these subtasks can be decomposed further into more subtasks. The process continues until the composite tasks are decomposed into primitive tasks.

Primitive tasks refer to the primitive actions –or cognitive steps– identified as useful for translators when trying word translations. The primitive tasks constitute the procedural knowledge (methods in the sense of object-orientation) associated to the entities of the dictionaries. For example, the *rths* (thesaurus-like search of concepts) action is one of the twenty five primitive tasks defined in the model. This action performs the search for lexical units, based on some constraints, in case the user has an imprecise idea about the exact concept s/he is looking for. The primitive task *rths* would be preferentially used when verifying the meaning of a source word or when finding a production hypothesis which are its parent-tasks.

The execution of the composite tasks is controlled at the strategic-level of the system, based on the information contained in the task-structure.

We will illustrate these ideas with the TARGET_WORD_GENERATION composite task (the last subtask to be performed to solve the main task) in figure 2. It is executed to get a suitable word that corresponds to the source word to be inserted in the context; in other words, its output is the corresponding target lexical word-form corresponding to a (pre-lexical) meaning, translation of the source lexical concept (figure 3). Some prerequisites must be fullfilled to activate this task: there is a source concept to be translated to an equivalent concept in the target language, the meaning and the morphological information associated to the source concept is well-known, the context in which it has appeared is also known; as well, the system knows some features about the translator.

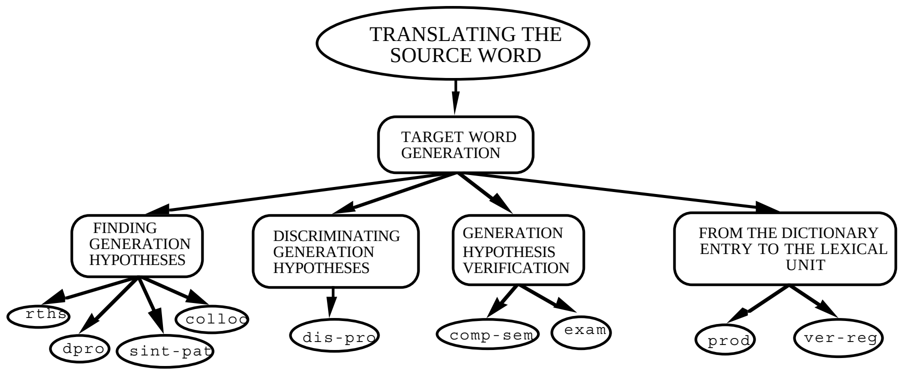

Fig. 3. Decomposition diagram of the TARGET_WORD_GENERATION task involved in the lexical translation process.

In our empirical model, four subtasks are involved in the resolution of the task: FINDING_GENERATION_HYPOTHESES whose goal is to produce the target word-form, given that the translator has enough knowledge about the translation of the source concept; DISCRIMINATING_GENERATION_HYPOTHESES which focuses on discriminating between the possible productions of the target concept by looking at their properties; GENERATION_HYPOTHESIS_VERIFICATION which verifies whether the final hypothesis for the target concept is correct in the context in which the source concept has appeared; and finally the composite subtask (FROM_THE_DICTIONARY_ENTRY_TO_THE_LEXICAL_UNIT) whose objective is to produce the final word-form which would be used as the translation of the source word-form . As can be seen in figure 3, these subtasks are expressed in terms of other subtasks, which in this case are primitive actions.

The description of the tasks is expressed in CML (Conceptual Modelling Language), which is one of the languages used in the CommonKADS methodology (Schreiber et al., 94). Each task, either composite or primitive, has a descriptive-frame in the task-structure. The TARGET_WORD_GENERATION task, for instance, is specified as follows:

```
task-knowledge
   task#36:  target_word_generation
      task-specification
         task-definition
            goal:           to get a word that
                            corresponds to the source word
                            word and is suitable to be
                            inserted in the context.
            input:          target_concept: concept
                            target_definition: definition
            output:         target_unit: lexical_unit
         task-body
```

```
type:           composite
parents:        translating_the_source_word
sub-tasks:      finding_generation_hypothesis,
                discriminating_generation_hypothesis,
                generation_hypothesis_verification,
                from_the_dictionary_entry_to_the_lexical_unit
additional-roles:
                context: text_model,
                translator: translator_model,
                languages: language*,
                list_of_productions: concept*,
                ?verified: boolean,
                target_entry: dictionary_entry,
                target_morphology: morphology
acceptance-conditions [3]:
                (available_concept(target_concept)) or
                (available_definition(target_definition))
competence-conditions [4]:
                (produced_unit(target_unit))
control-structure:
 -- algorithm which reflects the control
 -- to carry out the task
```

Primitive tasks are specified in a similar way. Below we present the description of the *rths (thesaurus-like search of concepts)* primitive task:

```
task-knowledge
   task#24:            rths
      task-specification
         task-definition
            goal:      to find the concepts that satisfy a given set of
                       constraints
            input:     constraint_expression: expression,
                       preferred_dictionary: dictionary,
                       result_language: language
            output:    list_of_concepts: concept*
         task-body
            type:      primitive
            parents:   meaning_verification,
                       finding_generation_hypothesis
            sub-tasks: nil
```

[3] The acceptance field states which conditions should be present in the working memory of the system for a task to be executable.

[4] The competence field states which conditions will be held in the working memory of the system, once a task has been executed.

```
additional-roles:
          result_concepts_variable: variable
```

3.3 An overview of the primitive tasks

As we have already mentioned, primitive tasks refer to the basic actions carried out by translators when using dictionaries. Hence, they can become procedural units that can be directly executed by the system.
These primitive tasks are classified according to the composite tasks where they occur.

3.3.1 Source text understanding

The definition request (*ddef*) is the core function in the word understanding task. It takes as input a concept, an explanatory-level, a dictionary and a language, giving a definition as output. The following example is a definition query for the meaning of guêpe (wasp) in the *Le Plus Petit Larousse* (LPPL) French dictionary, with inherited as the explanatory-level. The result is the textual definition along with other extra information ("inherited") that, not being explicit in the dictionary (in italics in the example), is deduced by MLDS.

`User.-` `ddef (|gu ê pe I 1|, inherited, LPPL, French, ?D)`
`System.-` `Wasp is an articulated hymenopterous insect`
with sting and legs;
A bumblebee is a wasp;
A wasp's nest has wasps.

Other functions related to this task are: reformulation of a definition (*rdef*), definition verification (*vdef*), request of properties of a concept (*dpro*), verification of properties of a concept (*vpro*), request of differences for two concepts (*ddif*), request of relationships between two concepts (*drap*), verification of relationships between two concepts (*vrap*), thesaurus-like search of concepts (*rths*), morphological analysis of a word form (*analy*) and request of examples (*exam*).

3.3.2 Searching for the equivalent

MLDS offers a set of primitive tasks in order to accomplish the complex task of searching for a suitable equivalent. The search for potential translation equivalents (*equiv*), the search for syntactic constructions that correspond to a given pattern (*synt_pat*) and the semantic compatibility between concepts according to a given relation (*comp_sem*) are the most relevant ones.

3.3.3 Target unit production

This task involves the thesaurus-like search of concepts (*rths*), the lexical collocation (*colloc*), the lexical form production (*prod*), and the verb-regime (*ver-reg*).

This is an example of the use of *rths*:

```
User.-   rths ((and (?X   HYPERONYME   |consumer I 1|)
                (?X   AGENT |feu I 1|)),
                 LPPL, Basque, ?X, ?LC)
         The user asks for verbs in Basque
         to consume with agent fire
System.- LC=(|izeki I 1|, |kiskali I 1|)
         to burn, to blacken.
```

4 The Dictionary Knowledge Base

The DKB is an important part of the MLDS System. The contents of the domain and inference layers have been represented in the DKB. A typed frame-based model has been used where concepts are represented by units that are interrelated by slots representing lexical-semantic relations. The proposed model allows us to represent different dictionaries of different languages. Each dictionary is treated in a uniform way and is represented as an enrichment of the representation used for monolingual dictionaries in previous works (Agirre et al. 1993a). The design of the DKB is based on the following principles:

A) Relational representation model. We extracted from the analysis of dictionary definitions different types of relations (Artola 1993) such us: synonymy and antonymy, taxonomic relations as hypernymy/hyponymy –obtained from definitions of type "genus et differentia"–, and taxonymy itself –expressed by means of specific relators such as *sorte de* and *espèce de*, belonging to the lexicographic description metalanguage–, meronymy, and others.

B) Typed information. Different types of objects have been defined and the proposed representation for them captures the common features which will be inherited. Fundamentally these types of objects represent attributes related to the lexical and semantic information. These descriptions constitute the meta-knowledge of the DKB.

C) Integration of multiple monolingual dictionaries. After studying different possibilities, our proposal considers that each monolingual dictionary defines its own separate conceptual world. However, between these separate worlds (dictionaries) there are common units which will be linked by means of equivalence relations.

D) Integration of multilingual dictionaries in a bilingual environment. Two models are considered as significant: Interlingual and Transfer. While the former proposes an intermediate representation for establishing the links between concepts or words from the source language to the target language, the latter establishes the links directly. In our system, word-senses are interlinked by means of complex transfer-relations. These relations allow us to represent different important phenomena: lexical gaps, translation of idioms, partial equivalence between word-senses, and so on.

E) Modularity. This aspect has guided us when deciding on the representation

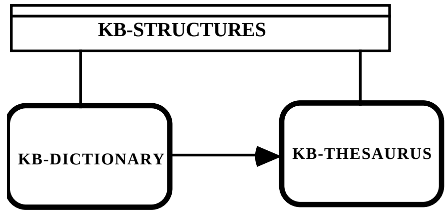

Fig. 4. The components of each Monolingual Environment.

and types of objects to be used or on the integration of different monolingual and bilingual dictionaries. In our case, monolingual components have been represented independently of each other. The bilingual environment puts them into correspondence.

F) Help in the translation process. Some aspects of the translation process itself must be taken into account in order to create a useful tool. Based on our experience, we have considered different types of information as useful: meaning, morphological, and syntactical information expressed by means of the following features: subcategory, idioms, collocations, and selectional restrictions.

G) Capability for making inferences. Using this capability, implicit information contained in the dictionary is made explicit.

4.1 Architecture of the Knowledge Base

The knowledge base of MLDS has been built to support the information included in two Basque and French dictionaries, but it offers a general framework for different languages. In the present version, the system is composed of four main knowledge bases: SDMOL1 and SDMOL2 contain monolingual information concerning the source and target languages respectively, whereas SBL1/2 links concepts from SDMOL1 with concepts in SDMOL2. The meta-knowledge level about basic objects used in all these DKB's is included in the S/STRUCTURES placed on the top level of the hierarchy (for a general view, see figure 6 in section 4.3).

4.2 Monolingual Environment

The monolingual modules (SDMOL1 and SDMOL2) are composed of the following three KB's (see figure 4):

KB-THESAURUS is the representation of the dictionary as a semantic network of frames, where each frame represents a *one-word concept* (word sense) or a *phrasal concept*. Phrasal concepts represent phrase structures associated with the occurrence of concepts in meaning definitions in the KB-DICTIONARY. Frames are interrelated by slots representing lexical-semantic relations to other frames such as synonymy, taxonomic relations (hypernymy, hyponymy), meronymic relations and specific relations realised by means of meta-linguistic relators. Other slots contain phrasal, meta-linguistic, and general information.

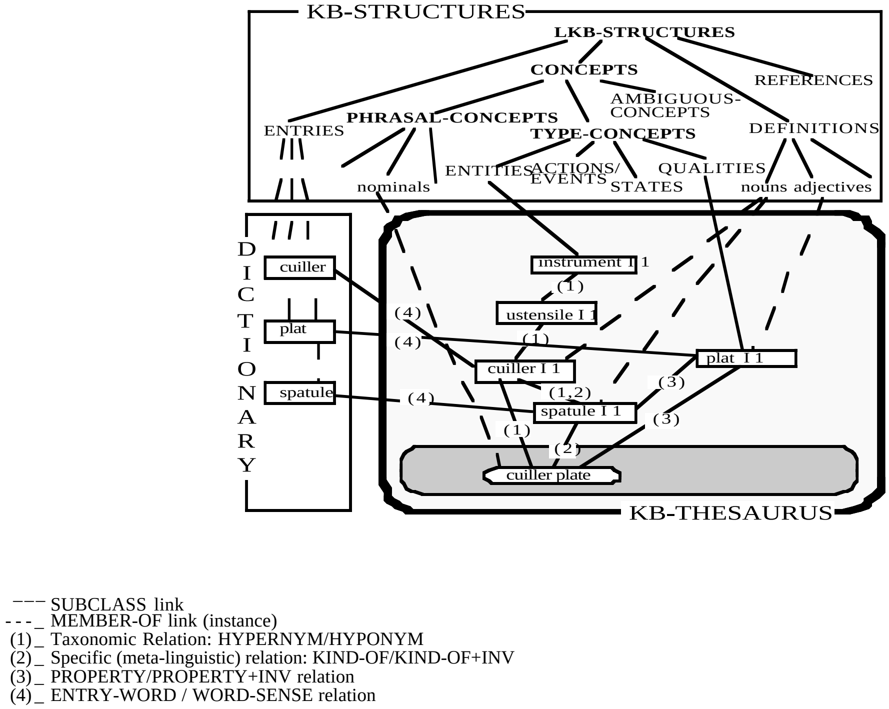

Fig. 5. The French monolingual Dictionary Knowledge Base.

KB-DICTIONARY represents the entries (words) of the dictionary; they are directly linked to their corresponding senses in the KB-THESAURUS.

KB-STRUCTURES contains meta-knowledge about concepts and relations described in KB-DICTIONARY and KB-THESAURUS: value ranges, inheritance modes, etc.

Figure 5 gives a partial view of the three knowledge bases which form the DKB with their corresponding units and their inter/intra relationships.

In the KB-THESAURUS, some of the links representing lexical-semantic relations are created when building the initial version of the knowledge base, while others are deduced later by means of specially conceived deduction mechanisms (Agirre et al. 1994c). When a dictionary entry like |*spatule I 1*|*: sorte de cuiller plate* (a kind of flat spoon) is processed, new concept units are created in KB-THESAURUS (and in KB-DICTIONARY) and linked to others previously included in it. Due to the effect of these links new values for some properties are propagated through the resulting taxonomy.

Figure 5 also shows the types of concepts used: |*spatule I 1*| and |*cuiller I 1*| are noun definitions and considered subclasses of ENTITIES while |*plat I 1*| (an adjective) is a subclass of QUALITIES. The phrasal concept unit representing the noun phrase *cuiller plate* is treated as a hyponym of its nuclear concept (|*cuiller I 1*|).

Our representation model allows the integration of different monolingual dictionaries. This is a hard task considering that, in some sense, each dictionary defines its own conceptual world. Nevertheless, equivalence relations among the concepts represented in the dictionaries have been established and represented. Knowing the difficulty of mapping among concepts from different dictionaries, only those links which map clearly equivalent concepts are proposed.

4.3 Bilingual Environment

In the bilingual environment, SBL1/2, two monolingual submodules are related by means of a Bilingual Module. The monolingual modules have been described above, here we describe the bilingual module in which links among the corresponding concepts of each monolingual environment are established. Note that the definitions of concepts of the monolingual environments do not change when they are integrated into the bilingual environment.

KB- S/Bilingual: This module includes the definition of the classes and attributes needed in the representation of the Bilingual Dictionary. Three different classes have been defined:

A) Source-Unit Class: Defines the type of the link between the unit of the Bilingual Dictionary and its corresponding concept of the Monolingual Dictionary. It is represented by the equivalence relation.

B) Target-Unit Class: The same as the previous but referring to target-units.

C) Bilingual Unit Class: Characterises the meaning information about the equivalence relation itself. Different attributes are used in this model in order to enrich the bilingual links between concepts of different languages. Some of them follow:

- level-of-equivalence: Represents the level of equivalence between the linked units. Three different levels of equivalence are proposed to characterise these links. One level represents that the concept of the source language is more general than its corresponding concept in the dictionary of the target language. A second level represents greater specificity, and the third level expresses conceptual equivalence.
- type of equivalence: The model of representation proposed permits us to establish two types of equivalence. Those which relate two concepts are named conceptual and those which relate concepts with phrasal concepts are named syntagmatic.

In the KB-S/BILINGUAL module (in our case KB-Basque/French) the information contained in our Basque/French Bilingual Dictionary is represented.

When a bilingual dictionary entry like *spatule: a kind of flat spoon* is processed, a new unit is created and the corresponding links with the monolingual dictionaries (source and target) are established by means of the EQUIVALENCE and the REPRESENTATIVE+INV relations.

An example of the bilingual environment follows (see figure 6). This figure shows

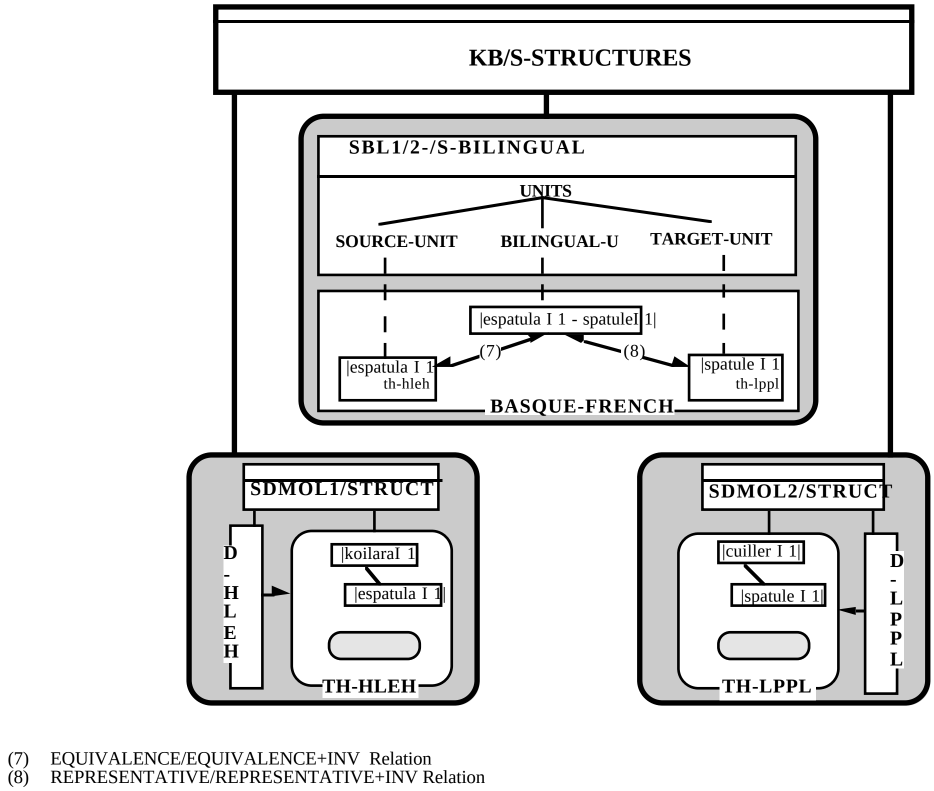

Fig. 6. General view of the Knowledge Base of the system.

the general organisation of the global DKB in which the correspondences between the monolingual and bilingual environments are established.

4.4 Translation-oriented Knowledge

As a result of the study about the use of dictionaries performed with human translators, the need for including other types of knowledge in our system has been detected. In this group we include morphological, lexical and syntactic information. In our system, the treatment of each of these types of knowledge implies the addition of new attributes and relations.

4.5 Size and contents of the DKB

A prototype of MLDS has been built. The size of both monolingual and bilingual environments of the MLDS prototype is explained next.

- Monolingual environments
  - A) Basque Dictionary Knowledge Base: It contains 168 entries, each one representing one word. The KB-THESAURUS contains 305 conceptual units;

89 in the DEFINITIONS class, 90 in REFERENCES and, 126 corresponding to phrasal concepts.

B) French Dictionary Knowledge Base: the knowledge base corresponding to 2607 entries of the LPPL has been created semi-automatically (Artola 1993). A subset of the Knowledge base was selected: It contains 541 entries, each one representing one word. KB-THESAURUS contains 1139 conceptual units; 418 in the DEFINITIONS class, 513 in REFERENCES, and 208 corresponding to phrasal concepts.

- Bilingual environment
  Our KB-BASQUE/FRENCH contains 158 Units in the SOURCE Class, 190 included in the TARGET Class, and 208 BASQUE/FRENCH units.

5 Functional Vision of MLDS

A computerised system with abilities to translate words could be used as an autonomous tool with a limited functionality. Basically, such a tool would translate a given word from one language into another. However, we prefer the MLDS to be human assistance-oriented and integrated into a more general framework. Namely, we are interested in endowing the MLDS with the means of interacting with the human translator. This interaction will enrich the performance of the human translator if it preserves some rules:

- The translator must be responsible for the translation of words. The dictionary system will help him/her, but decisions must be made by the human user. MLDS does not play the role of the translator.
- MLDS ought to show the real complexity of the task, warning the translator about problems and risks, and proposing different ways of finding, choosing or verifying. From our view, a system that helps by this process is much better than one which merely gives a proposal for a translation. The former system supports a decision without making it, the latter makes a decision without supporting it.
- The behaviour of the MLDS has to be transparent, nothing is to be hidden from the user. This rule is closely related to the relevance of the process itself.
- MLDS must adapt to the state of the translation when interacting. In a translation-state, conditions of the translator, of the languages, or referring to the objects and tasks can appear. It is particularly relevant to take into account the context of the word to be translated. MLDS will use the context when executing help-functions.
- MLDS should demonstrate intelligent behaviour, exploiting its deduction and anticipation capabilities.
- MLDS must not be annoying. We completely agree with (Hutchins and Somers 1992) in 153-154: "The difficulty with this kind of interactive system is that there are often so many interactions for a single sentence that it might have been quicker for the user to translate from scratch. Furthermore, it is frustrating and irritating to have constant repetitions of the same interactions with the same lexical items".

Therefore, the system ought to avoid interfering. From a "non obtrusive point of view" this is a central quality feature of an intelligent help system (Krause et al. 1993).

- MLDS should combine active and passive modes of aiding the user, by answering questions or queries as well as by giving unsolicited advice.

The system architecture shown in section 2 allows preservation of these rules by means of different strategies that are applied depending on the context in which the translation is carried out.

5.1 Use Strategies

The system can be roughly seen as a help system in the way shown by, for example, (Fischer et al. 1984; Breuker 1990). Nevertheless, it has extra features to maintain fluent and enriched interaction with the translator.

Three types of help-strategies have been designed.

- Question-answer strategy. This is the strategy followed when the user poses questions, i.e. passive help as in traditional systems. The translator requires information from the dictionary system by means of direct queries. There is a correspondence between these queries and the primitive tasks of the conceptual model. Each query activates a primitive function of the inference layer in the dictionary knowledge base. The task-trigger is supposed to update the working-memory in order to register the state of the translation.
- Feedback interacting strategy. In this strategy, as in the question-answer one, it is the translator who consults the dictionary-system by means of basic functions, but in this case the complexity of the function leads the system to ask for help from the user. Usually these kind of dialogues are due to the need for binding or disambiguating. The user could answer the system in order to improve its performance, however, answering is optional and users are not forced to do it. The dialogue is guided by the interface-manager.
- Autonomous behaviour. MLDS decides by itself to help the user, supposing that s/he needs it. This is what is known as active help in the literature (Breuker 1990). Typically, active help systems assist the user when error situations are produced. However, in our case it is not easy to detect when erroneous dictionary-queries are being made. This is why, instead of adopting this approach of active assistance, we have designed the autonomous behaviour as a way of giving complementary information whenever possible.

  Such an autonomous process is activated when the system detects that a query-sequence has been interrupted without concluding the translation. In this situation the MLDS presumes that the translator does not know how to follow and decides to use the knowledge about the translation process it holds.

  The complementary information is displayed in a specific window without interfering in the translators activity. If the translator considers that the complementary information is interesting, s/he uses it, if not, s/he can ignore it.

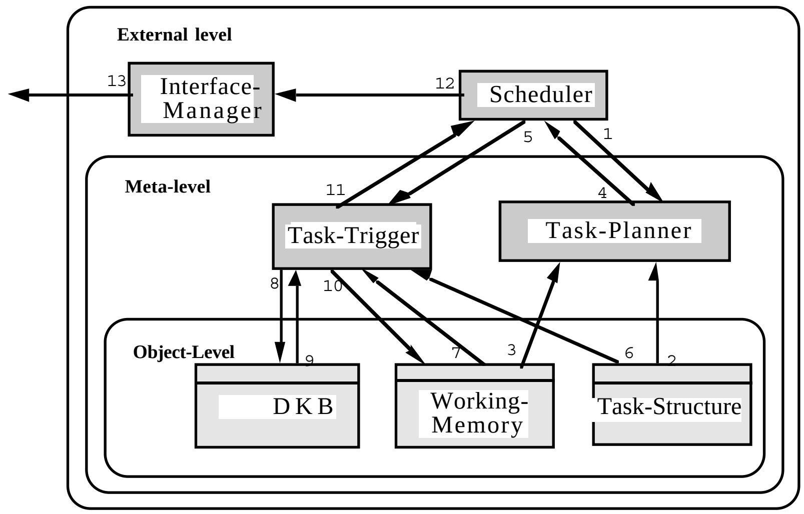

Fig. 7. Running of the autonomous process.

Following, we present the algorithm for controlling the running of this strategy:

1. The scheduler decides to start the autonomous processing, and passes the control to the task-planner (step 1 in figure 7).
2. The task-planner determines which is the most suitable primitive task to be activated by consulting the working-memory and some slots of the task-structure (namely the acceptance-conditions and the control-structure). It corresponds to steps 2 and 3 of figure 7.
3. The planner informs the trigger about the chosen task. Steps 4 and 5.
4. The trigger carries out the task updating the working-memory (steps 6, 7, 8, 9, and 10).
5. The interface manager shows the retrieved information to the translator in an auxiliary window (steps 11, 12, and 13).

In figure 7 these 5 stages are unfolded into 13 steps.

Steps 2 to 13 are repeated until the autonomous execution is ended. This process could be interrupted either by finding a complete satisfactory lexical translation, or when continuation is not possible, or when the translator changes the translation state. Obviously, during the autonomous process no feedback interaction is possible.

6 An Example of Using MLDS

As explained above, our system works in different ways, so that it has several strategies for interacting with translators. Here we will illustrate these strategies by means of the evolution of a session. For each strategy we will make a brief description, give the context in which the interaction takes place, describe the interaction itself and provide some remarks to clarify different aspects of the treatment.

6.1 First Strategy: Question-Answer Strategy

The system answers the question posed by the user.

Context

Let us take the basque word "etxaldean"[5] as the source unit to be translated from Basque into French.

The translator knows that the lemma of this word is "etxalde" and decides to pose a query to MLDS in order to clarify its meaning. MLDS will answer the query. This situation seems similar to a look-up in a dictionary but it is interesting to remark that the information given to the user is enriched by other information inherited from the representation and obtained by means of an inference process, as explained in (Agirre et al. 1994c).

Interaction

The set of queries is previously defined and corresponds to the primitive tasks.

In this case, let us suppose that the user wants to know the definition of "etxalde". S/he will use the *ddef* (definition request) primitive task.

```
User.-  ddef (|etxalde I 1|, inherited, LPPL, Basque)
MLDS.-  Etxalde: Nekazaritzako lurrak dituen etxea,
        familia bat bizi dena
"Etxalde" is a house with lands used for farming, where a
family lives
```

Remarks

Users have some graphic facilities, both for questioning and for receiving the given information. The answer is the text found in the dictionary along with complementary information that has been deduced. Inference abilities are exploited in this looking-up process.

6.2 Second Strategy: Feedback Interacting Strategy

Some types of questions do not have a simple resolution. Answering these questions involves decision-making. In these cases, information from the user could be helpful, so the system has to operate on a collaborative task.

Context

Let us suppose that after reading the answer given in the previous interaction and, due to the fact that the definition of "etxalde" suggests another Basque word ("baserri") to the translator, the user wants to know the differences between both Basque words.

Interaction

Disambiguation processes are commonly involved in this kind of interaction.

By means of *ddif* querying it is possible to ask what the differences between "etxalde" and "baserri" are (two words the translator cannot differentiate).

[5] etxaldean stands for "in the house environment" or "in the farm".

```
User.-   ddif (|etxalde| |baserri|)
Which are the differences between the Basque words
``etxalde'' and ``baserri''?
```

When the system tries to answer this question, it finds that one of the terms in the question is ambiguous and that it can have two possible meanings. In order to give a final answer this ambiguity must be resolved between the two possible meanings and the appropiate one chosen. The user is asked about this.

Different options are presented and the user is asked to select one.

```
MLDS.-   baserri  has two meanings.
         Please choose one of them or continue if you want an
         automatic resolution:
              |baserri I 1| [6]  : ^ lurraldea[7]
              |baserri I 2|: ^ etxea[8]
```

The system will make the choice given by the user, if so, and go on with the initial task. In case the user doesn't give any answer the system will make it automatically from the options (the first probably), and it will continue. In the latter case, the quality of the answer cannot be guaranteed.

For our purpose, we will suppose that the user selects the second option: |baserri I 2|. Once the selection has been made, the system is capable of giving an answer to the former question.

```
MLDS.-       |baserri I 2|: lurrak dituena
             |etxalde I 1|: lurrak dituena
```

No differences are detected. It is noted that both |baserri I 2| and |etxalde I 1| have lands.

Remarks

This interaction does not interfere with the activities of the human translator. Interacting with MLDS is just optional.

6.3 Third Strategy: Autonomous Behaviour

Context

When no explicit questions are presented to the system three possible situations may be occurring:

- *Situation 1:* The translator has already found a –probable– satisfactory translation. In this case complementary information could be useful for verifying the translation or, on the contrary, it might cause the translation to be changed because it shows hidden meanings or unexpected usage-features.

[6] |baserri I 1| and |baserri I 2| represent the two meanings associated to baserri and ˆ indicates that the term that follows is more general than the concept on the left: |baserri I 1|. So, in this case, both meanings |baserri I 1| and |baserri I 2| are expressed in terms of more general concepts.

[7] |baserri I 1|: farm area.

[8] |baserri I 2|: farm house

- *Situation 2:* Although the translator does not detect it, actually there is usable information in the dictionary that could be helpful. The trouble does not lie in the dictionary, but in the strategy of using it. This kind of obstacle does not have to be undervalued (Nuccorini 1994). With such an hypothesis it is expected that the autonomous behaviour will help.
- *Situation 3:* The dictionary has no information about the required translation. Any attempt to help based on the dictionary will be a failure.

Without really knowing which of the three described situations occurs, the system begins its autonomous process based on the contents of the working-memory where the available information, obtained from the user or automatically inferred, is stored.

```
        {Partial view of the working-memory:
              (user-model
                   (preferred-language Basque)
                   (preferred-dictionary HLEH)
                   (source-language-knowledge good)
                   (target-language-knowledge good)
                   (expertise little))
              (text-model
                   (type book-for-children)
                   (left-context <....>)
                   (right-context <....>))
              (basic-roles
                   (source-language Basque)
                   (target-language French)
                   (source-unit etxaldean)
                   (source-entry etxalde)
                   (source-concept etxalde I 1)
                   (source-definition "....")
                   (source-form-morphology (noun, inessive, singular))
                   (target-concept ?)
                   (target-unit ?))
}
```

At this point, as no questions are posed by the user, the scheduler passes control to the task-planner. This module consults the working-memory, that is, storage of the state of the translation, and the task-structure, that is knowledge about the lexical translation process, and decides which task is the most suitable to be activated.

Interaction

In the following example, the choosen task is "to-look-for-equivalent". The trigger executes the task and the result of the execution will be presented on the interface-window without directly interfering with the activity of the translator.

```
   User.-   LOOK-FOR-EQUIVALENT (|etxaldean|)
```

When the system tries to answer this question, it finds that there can be three French concept equivalents to the one of question. The three concepts are presented to the user.

The results of "look-for-equivalent" task is presented in the window and the user can chose one to use it or not.

```
MLDS.-   In this example, three French concepts are presented:
         ((|ferme I 4| [9] 0.5)
          (|b â timent I 10| [10] 0)
          (|maison I 1|[11] 0))
```

The system needs only one of them to go on with the autonomous process. It takes the first option and continues selecting another task. In this situation the definition request of |ferme I 4| is automatically selected by MLDS.

Similar to what happened in the previous step, the produced output is shown and the system goes on producing new hints.

The definiton of |ferme I 4| is shown in French.

```
MLDS.- (AND (|ferme I 4| HYPERONYME |domaine I 1| [12] )
            (|ferme I 4| ETAT |affermer I 1#1| [13])
```

This sense of ferme stands for a rented farm.

In an advanced step of the autonomous process a translation of the source unit will be proposed.

```
MLDS.-  "etxaldean"  translated as  "dans la ferme" [14]
```

Remarks

The last translation is achieved by integrating grammar functions into the system, that is, morphological analysis and generation that work as primitive tasks.

During this process the system may provide complementary help by means of examples, additional information and so on. The offered set of hints may assist the translator in the final decision.

7 Conclusions

We have presented MLDS, a computational dictionary system for human translators. This system was conceived to assist human users in the task of translating lexical units. In this presentation we have focused on the design-structure and the functional behaviour of the system, rather than on implementation details.

From a methodological point of view, it is innovative that the design of MLDS is based on the analysis of real human behaviour instead of a given theory. A model of the use of dictionaries when translating has been defined, and this model, written in a formal task-specification language, has been adapted to be executable.

[9] |ferme I 4|: farm
[10] |bâtiment I 10|: building, edifice
[11] |maison I 1|: house
[12] |domaine I 1|: domain
[13] |affermer I 1#1|: to rent
[14] "dans la ferme" French for "in the farm"

The specification and classification of the different types of knowledge as well as the design model of the system has been made using the principles of the KADS knowledge-engineering methodology.

The Dictionary Knowledge Base (DKB) is worthy of a quite detailed description. A frame-based representation model for dealing with multilingual lexical knowledge has been proposed. The basis of such a model was established previously for a monolingual knowledge base (Agirre et al. 1996). Upon this basis some enrichments such as the treatment of the multilinguality, or the incorporation of specifical translation-oriented knowledge has been introduced.

Concerning its functionality, MLDS works as an active help system. It is designed to be adaptable (it takes into account the state of the translation process), anticipative (it tries to anticipate the user' queries), and intelligent (with deduction abilities).

A prototype of MLDS has been implemented on a SUN machine using the KEE knowledge engineering environment. The future goal of this research is to implement a friendly interface and enhance the prototype allowing us to test, validate and evaluate the adopted model with human translators.

8 Acknowledgements

This research was supported by the Basque Government, and the University of the Basque Country in the projects EX98/39 and UPV 141.226-TA073/96.

We would like to thank Professor Jon Patrick from the University of Sidney for his useful comments.

References

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K.. Analysing word-level translation activity to design a computerised dictionary help system, (Panel Session) *EURALEX94* Congress, Amsterdam. 1994a.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K.. IDHS, MLDS: Towards Dictionary Help Systems for Human Users. *Semantics and Pragmatics of Natural Language: Logical and Computational Aspects* Sara, 167-188, 1993a.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K.. Sistema diccionarial multilinge: aproximación funcional. *Boletín SEPLN*, **14**, 313-333, 1993b.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K.. A methodology for the extraction of semantic knowledge from dictionaries using phrasal patterns, *Proceedings IBERAMIA'94*. Caracas. 263-270, 1994b.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K.. Lexical Knowledge Representation in an Intelligent Dictionary Help System, *Proceedings COLING'94*. Kyoto. 544-550, 1994c.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Soroa, A.. Constructing and intelligent dictionary help system. *Natural Language Engineering.* **2** (3), 229-252. Cambridge University Press, 1996.

Arregi X.. ANHITZ: Itzulpenean laguntzeko hiztegi-sistema eleanitza /Sistema diccionarial multilinge de ayuda en la traducción, Ph.D. Thesis. UPV-EHU. 1995.

Artola X., Evrard F.. Dictionnaire intelligente d'aide la compréhension. *Actas IV Congreso International EURALEX'90*. Benalmádena. 45-57. Barcelona: Bibliograf, 1992.

Artola X.. HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception et construction dun systme intelligent daide dictionnariale (SIAD), Ph.D. Thesis. UPV-EHU. 1993.

Atkins B.T., Knowles F.E.. Interim report on the EURALEX/AILA Research Project into Dictionary use, in T. Magay, J. Zigány eds., 381-392. *Proceedings BudaLEX88.* Budapest: akadémiai kiadó, 1990.

Breuker J. (ed.). EUROHELP. Developing Intelligent Help Systems. Report on the P280 ESPRIT Project EUROHELP. Kopenhague. 1990.

Cole R., Mariani J., Uszkoreit H., Varile G.B., Zaenen A., Zampolli A., Zue V. (eds.). *Survey of the state of the art in human language technology* **8**. Studies in Natural Language Processing. Cambridge: Cambridge University Press. 1997.

Dagan I., Church K.W. Termight: Identifying and translating technical terminology. *Proceedings of the 4th Conference on Applied Natural Language Processing.* 34-40. Stuttgart, Germany. 1994.

Fillmore C.J., Atkins T.S.. Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography, in B.T.S. Atkins, A. Zampolli eds., 349-393. *Computational Approaches to the Lexicon.* New York: Oxford University Press, 1994.

Fischer G., Lemke A., Schwab T.. Active Help Systems, in T. Green, M. Tauber, G. van der Veer eds., 116-131. *Proceedings of 2nd Conf. on Cognitive Ergonomics.* Heidelberg: Springer Verlag, 1984.

Hartman R.R.K.. Four perspectives on dictionary use: a critical review of research methods. The Dictionary and the Language Learner, in A.Cowie ed., 11-28. *Papers from the EURALEX Seminar.* Leeds. Lexicographica, Series maior, 17, Tbingen; Niemeyer. 1985.

Hatherall G.. Studying dictionary use: some findings and proposals, Proceedings LEXeter'83, 183-189, Lexicographica, Series maior 1, Tbingen; Niemeyer. 1984.

Hauenschild C.. The Lexicon and the General Design of MT Systems, Machine Translation and the Lexicon, Workshop of the European Association for Machine Translation, Heidelberg, 1993.

Hutchins W.J., Somers H.L. *An Introduction to Machine Translation.* Cambridge: Academic Press, 1992.

Isabelle P.. Bi-Textual Aids for Translators. *Proc. 8th Annual Conf. of the UW Centre for the New OED and Text Research.* Waterloo. 76-89, 1992.

Johnson R.L., Whitelock P.. Machine translation as an expert task, in S. Nirenburg ed., 136-144. *Machine Translation: Theoretical and methodological issues.* Cambridge: Cambridge Univ. Press, 1987.

Kay M.. The Proper Place of Men and Machines in Translation, CSL-80-11. Xerox PARC, 1980.

Krause J., Mittermaier E., Hirschmann A.. The Intelligent Help System COMFOHELP. Towards a Solution of the Practicality Problem for User Modeling and Adaptive Systems. *User Modeling and User-Adapted Interaction* **3d**. 249-282, 1993.

Martin W. and Al B.P.F.. User-orientation in dictionaries: 9 propositions, in T. Magay, J. Zigány eds., 393-399. *Proceedings BudaLEX88.* Budapest: akadémiai kiadó, 1990.

Neubach A. and Cohen A.D.. Processing Strategies and Problems Encountered in the use of Dictionaries, Dictionaries. *Journal of the Dictionary Society of North America* **10**. 1-20, 1988.

Nuccorini S.. On Dictionary Misuse. *Proceedings EURALEX'94.* Amsterdam. 586-597. 1994.

Sager J.C.. *Language Engineering and Translation: Consequences of automation.* Benjamins Translation Lib., 1994.

Schreiber G., Wielinga B., Akkermans H., Van de Velde W.. CML: The CommonKADS Conceptual Modelling Language. *8th European Workshop EKAW'94.* 1994.

Schreiber G., Wielinga B., Breuker J. (eds.). KADS: A Principled Approach to Knowledge-Based System Development. *Knowledge-Based Systems* **11**. London: Academic Press, 1993.

Starren P. and Thelen M.. General dictionaries and students of translation: A report on the use of dictionaries in the translation process. in T. Magay, J. Zigány eds., 447-458. *Proceedings BudaLEX88*. Budapest: akadémiai kiadó, 1990.

Steiner E.. Producers - Users - Customers: Towards a Differentiated Evaluation of Research in Machine Translation. *Machine Translation* **7**. 281-284, 1993.

Dictionaries

Le Plus Petit Larousse. Paris: Librairie Larousse, 1980.

Aurrekoetxea G. et al.. *Euskararako Hiztegia*. Bilbo: Adorez, 1986.

Lhande P.. *Dictionaire Basque-Fran cais*. Paris: Gabriel Beauchesne ed., 1926.

Sarasola, I. *Hauta-lanerako Euskal Hiztegia*. Donostia: KUTXA, 1991.

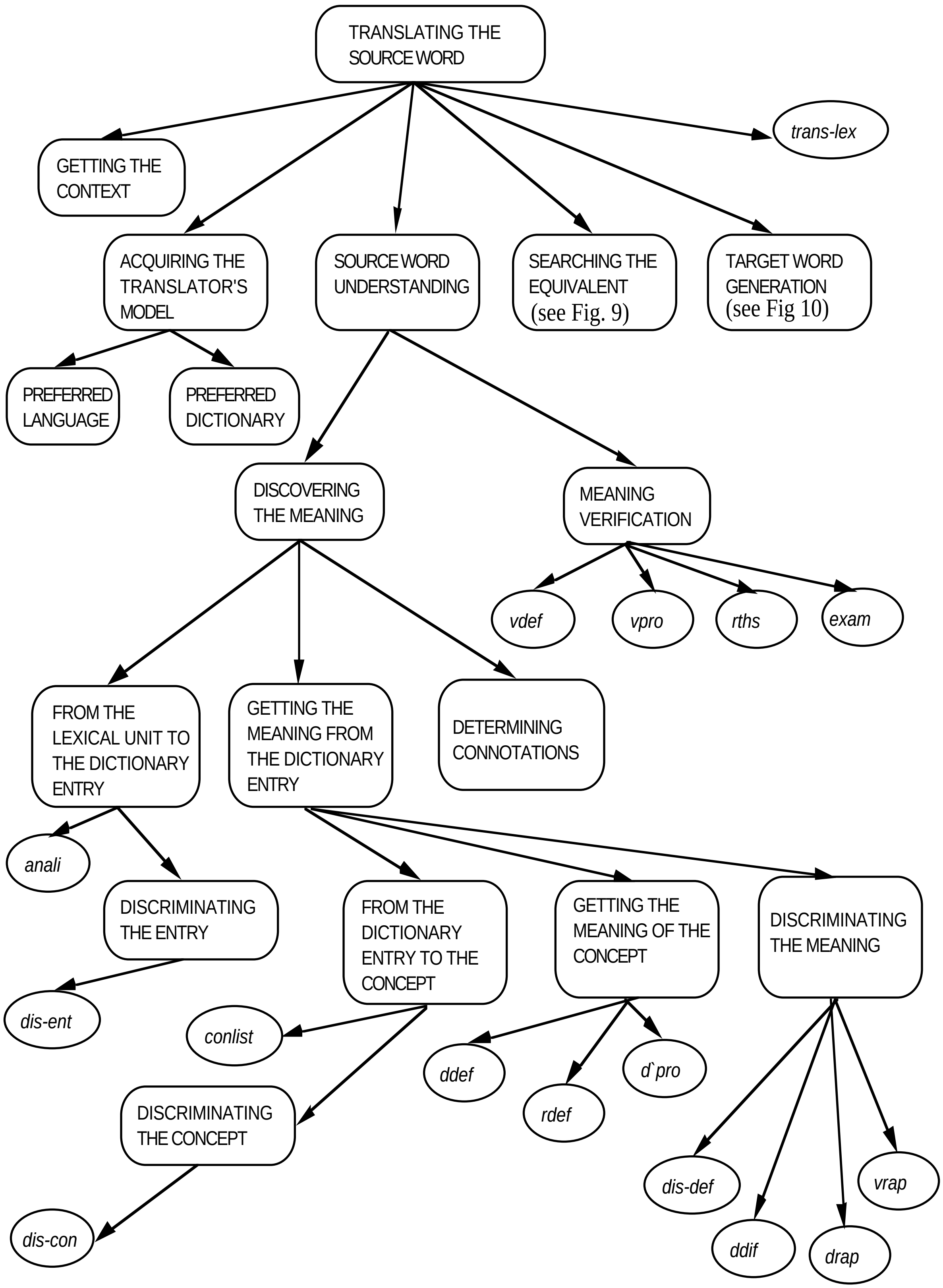

Fig. 8. Appendix 1: Decomposition diagram of the tasks involved in the lexical translation process (Part I)

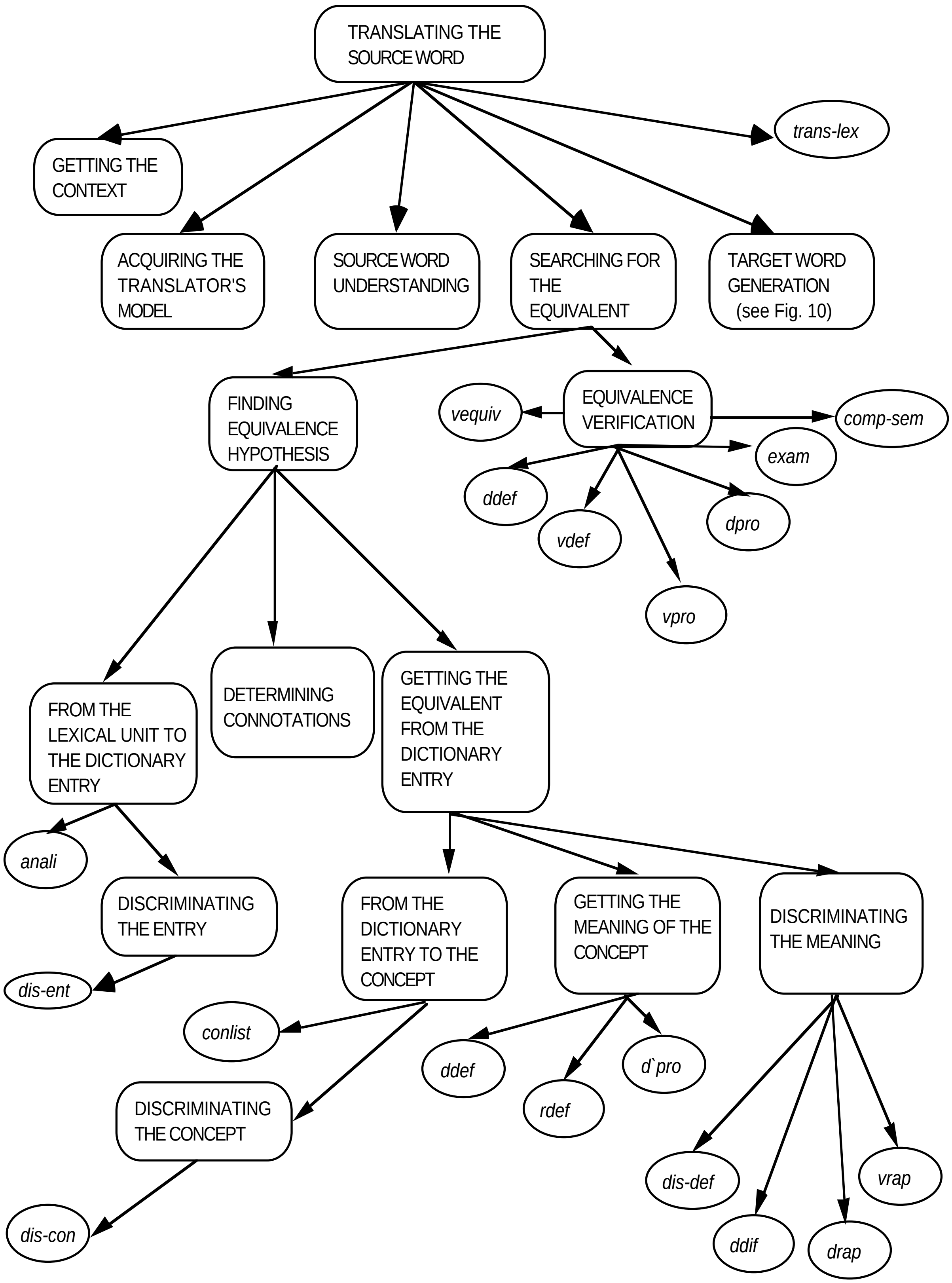

Fig. 9. Appendix 1: Decomposition diagram of the tasks involved in the lexical translation process (Part II)

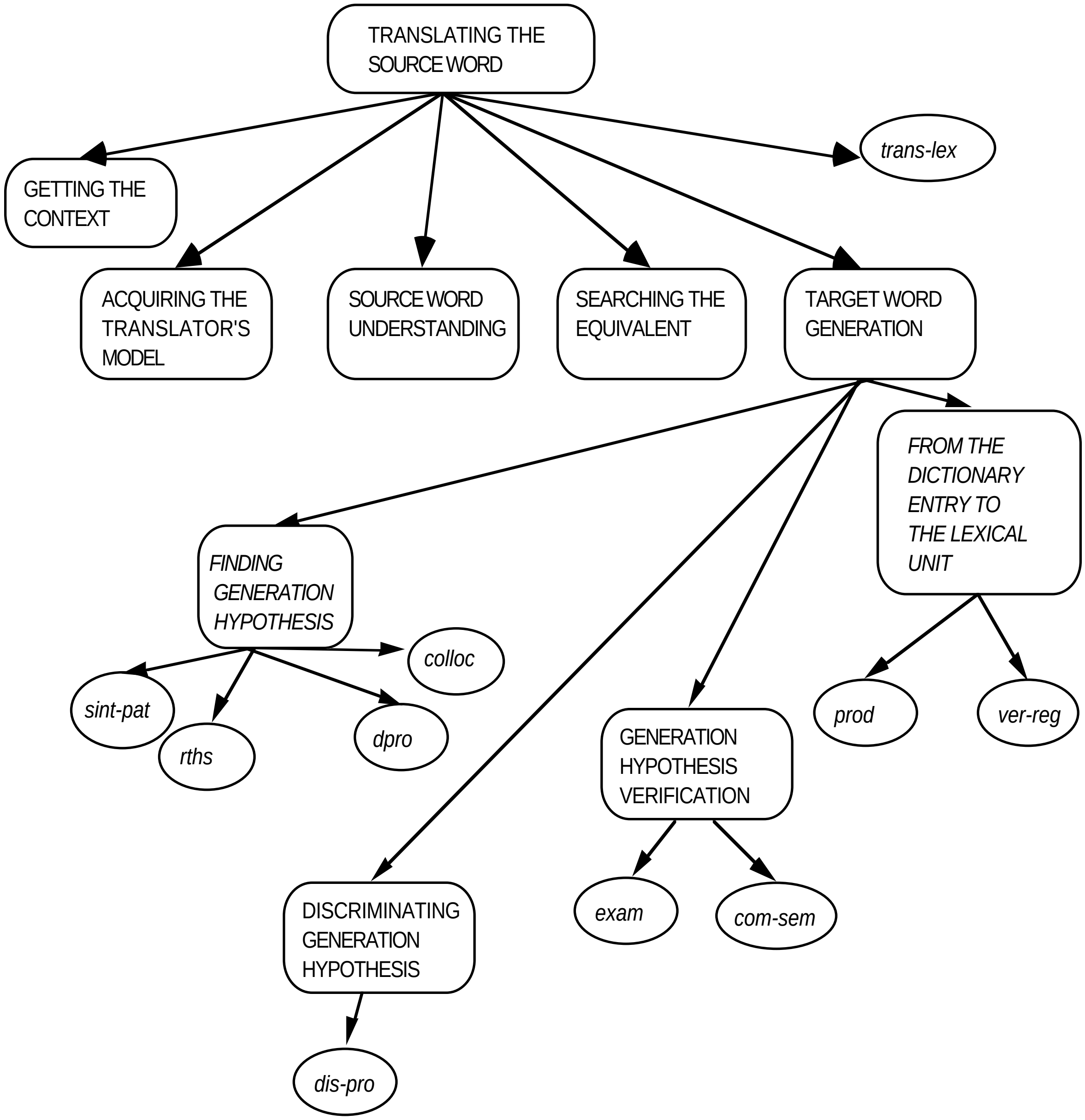

Fig. 10. Appendix 1: Decomposition diagram of the tasks involved in the lexical translation process (Part III)