EAGLES

Expert Advisory Group on
Language Engineering Standards

EAGLES

Lexicon architecture

EAGLES Document EAG–CLWG–LEXARCH/B

Version of Oct, 1993

Contents

1 Authors

Bruno Menon	
GSI ERLI	
Place des Marseillais, 1	
Charenton le Pont cedex	Tel: +33.1.48.93.81.21
France	Fax: +33.1.43.75.79.79
94227	E-mail: bruno.menon@erli.gsi.fr

Nicole Modiano	
CAP GEMINI INNOVATION	
86/90 rue Theirs	
Boulogne-Billancourt cedex	Tel: +33.1.49.10.52.05
France	Fax: +33.1.49.10.06.15
92513	E-mail: modiano@capsogeti.fr

2 General aims and objectives of the architecture work in EAGLES

- The objective of this subtask [1] is to propose a multilingual and multifunctional model for a dictionary viewed as a resource out of which to extract specific application lexicons. This implies among others that no need arises for
  - particular devices for redundancy-free representation, except where redundancy would be problematic for the linguistic description;
  - any bias towards a given type of storage;
  - any bias towards one type of application.
- EAGLES should provide a stable, robust, practical core dictionary layout which is extensible. This implies an ordering of the topics to be dealt with: the simple and relatively easily described facts should be treated in the core of the EAGLES recommendations, more problematic or controversial things in later extensions.

  Extensibility implies keeping track of possibilities of adding and/or (whole classes of) descriptive devices, e.g. lexical rules (in syntax, morphology, semantics, etc.).
- The EAGLES core architecture must be simple. It must be proven to work, to cater for the major relevant phenomena and to be extensible.

[1] Participants:

- Bruno Menon (ERLI),
- Nicole Modiano (CAP),
- Haroutioun Khatchadourian (CAP),
- Nicoletta Calzolari (PSA),
- Ulrich Heid (STR).

3 Input to the EAGLES architecture work: survey of Multilex and Genelex work

Multilex is the project on the lexicon which has the most explicit architectural proposal to make. Genelex has not produced a report on its architecture, but the existing implicit descriptions of the Genelex lexicon structure can be used as an input. Differences will be assessed in section 4. No architectural description of the lexicon for formal theories (e.g. HPSG, LFG) is available which would be as detailed as that of Multilex. However, HPSG's *sign-based* approach is described in the literature. DELIS, for example, will follow the "spirit" of the HPSG approach, yet without producing HPSG-syntax-like dictionaries.

3.1 What does "linguistic architecture" mean ?

- The linguistic architecture defines the basic objects of the model and their relations. It also specifies the general terminology which is common to the whole standard and used to talk about the dictionaries, their components and the interaction of these.
- If the lists of features for each linguistic level and each language are not rigidly restricted (as is the case in Multilex), the linguistic architecture defines relations between blocks of features instead of between individual features.
- Because most of the words described in a dictionary are quite regular and can be accounted for within a consistent framework while a considerable number of words are quite complex and variable, the linguistic architecture defines a robust basic structure and provides the means of building the additional apparatus when required.
- Finally, the linguistic architecture guarantees that the model is consistent at the top level whatever the language. This idea is very important: in database technology, a distinction between meta-schemata, schemata and instances of schemata is made. The meta-schema defines general well-formedness conditions for schemata. Schemata then define the logical format of language-specific and level-wise linguistic descriptions. Instances are then individual lexicons for which there is a translation relation expressed between the individual format of the instance and the "type" defined by the schema.

  The linguistic architecture should in our view make clear

  - at which level the standard operates: meta-schema, schemata or instances; we assume, as a working theory, that the standard will define the metaschema and the general parts of the schemata, not necessarily all of the language specific schemata, and certainly not the instances;
  - which elements and components of the schemata are needed for the description of different linguistic levels.

3.2 An overview of the Multilex architecture

- The Multilex model was designed from the start as multilingual. This viewpoint has motivated important choices such as

- flexibility of the model which should account for all the languages of the European community,
- identification of information nodes to allow for language equivalences,
- language-independant overall architecture.
- As a multifunctional model, MULTILEX can integrate information of a variety of theories and information intended to many applications (text analysis, automatic or computer-aided translation, etc.)
- The MULTILEX model is described as an SGML DTD (the MLEXd) for reasons of portability. Its design is independent from any internal database representation.

Global structure of MULTILEX dictionary A monolingual dictionary designed in accordance with the MULTILEX standard consists of:

3.2.0.1 Language (obligatory)

The name of the language which the entries of the dictionary belong to.

3.2.0.2 Language-specific information (optional)

The core of the standard is language independant. However, some features are not. Specific attributes have to be defined or inhibited for each language. For example, gender is not relevant for nouns nor adjectives in English but it is in Italian, French, Spanish, German, etc. This is also true for values. The values for gender in German are masculine, feminine or neuter while in French, Italian and Spanish, neuter is irrelevant.

Standardisation does not mean a frozen model. Each dictionary designer is free to define his/her sub-model with specific constraints providing that it complies with the standard. In this section it is possible to define local specific features:

- add values to the lists which are declared open,
- inhibate values in the lists declared free,
- add an attribute or sets of attributes for unforseen or insufficiently structured areas,
- change, when permitted, the status of an attribute or sets of attributes this status can be:
  - obligatory
  - inhibited
  - default = value
  - optional

3.2.0.3 Rule specifications (optional)

Rule specification which makes it possible to annotate the dictionary with grammatical information. It is often necessary to refer to a rule which is applicable to an item without having to describe this rule. A reference name is generally used. For example, plural formation rules imply a series of manipulations on the stem and the endings which are performed by the client NL application. An identifier of the rule is preferable to the actual name of the software module used in a specific application if the lexical resource is to be re-usable. An even better identification is to associate to this identifier the description of what the rule does in a simple and formalised way. This increases the readability of the data and its portability because it is possible to change the coding system for the rules without loosing functional description and without having to change the value names in the body of the dictionary.

3.2.0.4 Body The body of the dictionary consists of the entries, i.e., the objects allowed in the dictionary by the standard.

3.2.1 A two nodes model

The MULTILEX linguistic architecture is based on the independant description of, on the one hand, the words as they behave in the oral and written codes and, on the other hand, the meanings expressed by the words within a language. Lexical information is divided in accordance with this distinction and distributed around two nodes.

1. One node – called GPMU (Graphic-Phonologic-Morphological Unit) – is a triple identifying one set of coherent orthographic, phonological and morphological behaviours belonging to a canonical form.

   The canonical form is used in MULTILEX as an organising principle as one may speak of the different meanings of a word because a meaning is always associated to a form. The orthographic standard form is fixed in our languages by convention, apart from a number of exceptions (for example, *colineau* or *colinot* in French, or *yoghurt*, *yogurt* or *yoghourt* in English). In these cases, each form is covered by a distinct GPMU.

2. The other node – called LU (Lexical Unit) – identifies a single meaning within a language.

Lexical description results from the linking of the appropriate GPMUs and LUs. (see example: Annex number 1 “THE *BALLE* SCHEMA”)

- several LUs can share the same GPMU (this case accounts for homonymy)
- one LU can be linked to several GPMUs (this case accounts for variants)
- special behaviours cause a GPMU to be created (this case accounts for restricted usages).

3.2.2 Linguistic levels

The "blocks of features" mentionned above refer to the traditional "linguistic levels":

- orthography
- phonology
- morphology
- syntax (attributes relevant to each syntactic category, syntactic frames, selection restrictions, etc.)
- semantics (explicitation of the semantic content of a meaning)

Additional sets of information, which are not "levels" in a traditional sense are attached to these blocks when it is necessary. They consist of terminology pragmatics involving communication situations, professional usages, standardization and prescriptions, etc. and general pragmatics related to geographic preferences and diachronic specificities.

3.2.3 GPMU and CGPMU

3.2.3.1 GPMU The lexical architecture does not impose any hierarchy among the GPMUs. Differenciations between long forms, acronyms, abbreviations, variations, etc. are expressed by features. Also, preferences motivated by standardization or by pragmatic reasons can be attached to a GPMU without constraining the overall model. For example, "DNA" and "deoxyribonucleic acid" are two forms of the same meaning. They are described by two GPMUs each one with the appropriate feature indicating their type "initials" or "long form".

Variations like "cheque" and "check" are also dealt with by creating two GPMUs. In order to specify that the first orthography is British and the second is American, pragmatic information concerning the geographic usage is attached to each GPMU as a whole.

The GPMU can be either seen as the actual receptacle in which the orthographic, phonological and morphological information is stored or the node from which these constituents can be accessed with specific identifiers. This alternative does not change the role of the GPMU as the only surface object directly known to the LU. But mainly for facilitating implementation it has been decided that the GPMU constituents could be separated. The reason is that it may be easier to manipulate small objects than big ones. Also, this way, redundancy is dealt with straightforwardly.

3.2.3.2 CGPMU Complex words – compounds, phrases, expressions, etc. – are considered similar to simple words with regard to the two node organization. The word "Application Programming Interface" is described by one LU for the single meaning and two GPMUs ; one for the long form and one for the initials "API". The difference lies in the treatment of the long form, at the GPMU level.

In order to account for the internal structure of complex words, a special kind of GPMU, the CGPMU (Complex GPMU) has been created. This object is similar to the GPMU in many respects. It contains orthographic, phonological and morphological information but, in addition, it refers to the components – GPMUs of simple constituents – of the complex word, gives its morphological structure and global behaviour and expresses constraints on the components, if required.
For example, the CGPMU for "Application Programming Interface" refers to the three GPMUs "application", "programming" and "interface", and it also specifies that the mark of the plural only apply to "interface".
The content of a CGPMU can be summerized as follows:

- internal structure: description of the internal structure of the complex word with the number of components and their part of speech,
- orthography,
- morphology and morphological constraints: part of speech of the complex word, inflexional paradigm and information wether a component is inflected or not when the complex word is inflected,
- GPMUs of the components.

The CGPMU is an extension to the core MULTILEX standard. It was developed within a MULTILEX-EUROLANG [2] joint group.

3.2.4 The LU node

The LU is defined as the identifier of a single meaning. The meaning of "meaning" can be specified by a semantic theory (if there is one available) or it can refer to the intuitive notion used by traditional lexicography to separate the meanings listed under the canonical forms.
Blocks of features are associated to the LU (See Annex number two "THE MONOLINGUAL ARCHITECTURE SCHEMA"):

- one and only one semantic block,
- one or more syntactic blocks,
- zero or more bilingual transfer blocks.

3.2.4.1 Semantics Two different approaches to semantics have been proposed in the MULTILEX project:

- Logically regimented explanation of ordinary language entries from existing dictionaries (developed at the University of Bochum, Germany).

[2] EUROLANG is a EUREKA project aiming at creating a toolbox for Machine Translation applications.

- Metalinguistic constructs based on a set of semantic features (developed at the University of Pisa, Italy).

Both these theories are compatible with the MULTILEX standard, and they provide valuable guidance to dictionary designers. However, the objective is not to fix and impose a semantic approach, but to help identify, clarify and compare pieces of lexical information in order to facilitate lexicon design, usage and re-use. Choosing one semantic theory would exclude all – or nearly all – the others. Also, at present, there is no available multilingual lexical resource large enough to demonstrate universal validity of a single semantic theory. There is still an enormous amount of work to do before some uncontroversial results are achieved.

As for the core standard, at least a text definition, as in paper dictionaries, should be given as semantic information for human disambiguation purposes.

3.2.4.2 Syntax In most cases, one syntactic structure is related to a single meaning. Therefore for most of the LUs one syntactic description should be sufficient to cover the syntactic behaviour. However, variant syntactic structures may express exactly the same meaning. Straightforward cases can be dealt with without doubling the syntactic description block by operating at attribute and value level.

- Specific attributes
  For example, "to give someone something" is equivalent to "to give something to someone". Since this particularity concerns many verbs and since it is language dependant, a specific syntactic attribute is created to convey this information.

- Multi-valued attributes
  The syntactic variation often affects a single isolated element, e.g. a preposition. For example, a meaning of the French verb "comparer" can be expressed either by both the syntactic structures "comparer quelquechose *à* autre chose" and "comparer quelquechose *avec* autre chose". *à* can be substituted to *avec* and vice-versa without additional change to the structure. For unconstrained alternate variations, multi-valued attributes can be used. More than one attribute of the syntactic description can receive multi-values, if all the possible combinations of values are compatible.

When the variation cannot be expressed by these means, or if the design of a dictionary requires that each syntactic variation be represented in a single block, it is possible to define as many syntactic blocks as needed and to relate them to the same LU.

Examples of syntactic behaviours can be included within each syntactic block. It is highly recommended to use this possibility.

3.2.5 Cross-references

A cross-reference consists of:

- an implicit source, i.e., the object from which the link is created,

- a target which can be an identifier or a list of identifiers (LUs or GPMUs according to the source identifier),
- a type specifying the name of the relation established between the identifiers,
- an information on possible reversibility of the link,
- the name of the reverse link when it exists.

Generic links are proposed to the dictionary designer to be used at his/her convenience for expressing relations between LUs on the one hand, and GPMUs on the other hand.

These two cross-reference systems are clearly distinguished although they possess similar features:

- lu-cross-references allowing to relate LUs.
- gpmu-cross-references allowing to relate GMPUs. CGPMUs and GPMUs are not differenciated in cross-references. They are considered as similar.

In the core linguistic architecture, cross-references from or to individual blocks of features are not advocated because no really convincing evidence of their use has been produced. However, experimentation of the standard on a large number of entries may prove a special kind of cross-references to be of use between GPMUs and syntactic blocks when more than one syntactic block is attached to the same LU. In this case, specific sets of cross-references should be created and clearly identified in order to control possible side-effects especially when exploiting cross-references for extractions and export.

3.2.6 Terminology

MULTILEX chose not to give terms a distinct linguistic status from general language words. Both have meanings and surface behaviours which are described and related in the same manner. What makes a term LU different is that it has to be related to sub-language(s) and conceptual contexts to which it belongs. The term is considered from its linguistic sign side in the core linguistic architecture and dealt with as such although information concerning the concept side is refered to in the semantic block.

Terminological contexts of such and such sub-language show special constraints, restrictions or relaxations on the usage of terms. These behaviours may affect any linguistic level. Therefore, information on domains and sectors which characterize a sub-language layer and which specify the usage of the term are not concentrated in a terminology specific feature block. This information is attached to the affected feature block.

Sets of terminology relevant features – including domains, sectors, communication situation, etc. – can be attached to:

- GPMUs and CGPMUs,
- Syntactic blocks,
- Semantic block.

3.2.7 The multilingual architecture

The LU in the lingusitic architecture is seen as an operational notion. It identifies meanings within a language. In other words, the LUs of a particular language support and organize the semantics of this language. Although the LUs and their organizations are different in each language, the LU can be considered as an anchoring for multilingual relations. Equivalences between languages are obtained by drawing relations between their specific LUs according to relational rules explained hereafter.

Equivalence between LUs of different languages is multi-bilingual, transfer-based and unidirectional (See Annex number 3 “THE MULTILINGUAL ARCHITECTURE SCHEMA”).

3.2.7.1 Multi-bilingual Each dictionary in a lexical resource complying with the MULTILEX standard is basically monolingual. From each LU of each monolingual dictionary, it is possible to specify equivalences to LUs of several other monolingual dictionaries.
Example:

DE ”Fahrrad-lu...”	EN ”bicycle-lu...”
	NL ”fiets-lu...”
	FR ”bicyclette-lu...”

Each bilingual equivalence is dealt with separately. The features describing an equivalence for a given LU to an LU of another language is collected in a separate block and each block is attached to this LU. If several equivalents of the same LU exist in the other language, then more than one feature blocks can be created to account for each equivalence relation.
Example:

DE ”Fahrrad-lu...”	EN “bicycle-lu...”
	EN ”bike-lu...”
	EN ”cycle-lu...”
	NL ”fiets-lu...”
	NL ”rijwiel-lu...”
	FR ”bicyclette-lu...”
	FR ”vélo-lu...”

Conversely, when two LUs in language L1 has the same equivalent in language L2, then both the LUs of L1 point to the same LU of L2.
Example:

FR ”rivière-lu...”	
	EN ”river-lu...”
FR ”fleuve-lu...”	

3.2.7.2 Transfer-based The equivalence system is based on a transfer model. Each monolingual dictionary deals with the description of LUs according to monolingual criteria, and transfer provides the means to accommodate the source information to the target language needs. The notions of ”source” and ”target” dictionary are only valid from the point of view of transfer. Otherwise, each monolingual dictionary can be source and target. The bilingual equivalence block contains – along with the target LU – bilingual contrastive information, i.e., meaning discrepancies between the source and target LUs and conditions for syntactic matching. It also contain a characterization of the equivalence relation which can be *full*, *partial* in several ways or *nil* for different reasons.

Although LUs are semantic objects, the transfer is mainly syntax-based. This is not a restriction imposed by the standard but it results from the fact that large lexical resources do not provide extensive semantic descriptions yet which could supply enough information to perform semantic-based matching. The syntactic part of the standard provides the basic features from which to derive an interface format for bilingual syntactic matchings including elementary semantic features to apply selection restrictions.

3.2.7.3 Unidirectional For each bilingual transfer the source LU is considered from the point of view of the target language and transfer blocks are made up according to the target dictionary LUs. Note, however, that the source language description at the semantic level always remains unaltered. The fact that new transfer entries or target languages are added, never affects the source LU.
The testing material – used for disambiguating the meaning of the source LU against the target language – and the directions used for the accommodation of the source syntactic structure to the target syntactic structure apply to one language pair and one direction only.

The consequence is that the resulting transfer information is not reversible. Such an organization has been proposed, mainly for reasons of feasibility. There is also an organizational advantage to this structure: each language community, participating in a project involving multilingual data can have the organizational and managerial responsibility over one source language and one unidirectional language pair.

However, it should be clearly stated that the architecture deals with the linguistic and logical organization of the data, and that it leaves out problems relevant to applications. For example, the transfer blocks can be automatically reversed for a number of LUs marked as fully equivalent when specific translation lexicons are extracted from the main dictionary.

3.2.8 Conclusion

The MULTILEX linguistic architecture is meant to supply the backbone to the dictionary design. It defines the blocks of features and rules the type of relations they can bear. The relations restricted to the "containers" allow the flexibility of the content – the features –, the selection of which is a function of the needs of the languages involved, of the designers' projects and of the available data and expertise.

The MULTILEX linguistic architecture combines the lexicographic approach with the multidimentional facilities the electronic media offers. The organization of the data is allowed to provide the logical and linguistic ground for building user interfaces which would give access to linguistic facts in a way which is natural to lexicographers for the tasks they perform dayly. It also permits to safely and easily define sub-models derived from the initial linguistic architecture for extracting consistent sub-sets of the information from the main resource to a specific application lexicon.

3.3 An overview of the GENELEX architecture

Although Project Eureka GENELEX[3] includes tasks about conception, design and realization of pieces of software that will be used to manage, check and exchange lexical data, this section will only deal with its linguistic aspects.

3.3.1 General methodology

3.3.1.1 From pluralism of languages to multilingalism The general dynamics of project GENELEX is to design models for computational dictionaries of the languages it deals with in a separate manner (although in close cooperation). Once a sufficient level of specification has been obtained in these languages (i.e. for the time being French, Portuguese, Italian, Spanish, and, to a lesser extent, English), GENELEX will proceed with a search for general convergence and with a more precise definition of the multilingual dimension.

3.3.1.2 Generality of Format The issue involved in the development of GENELEX(GENEric LEXicon) is the generality of its format, which covers the following aspects:

1. maximum coverage: to take into account, for a given data entry, the maximum amount of non-redundant linguistic information. Information that can be deduced systematically from other linguistic information will not be recorded. In addition, the information to be recorded must not be limited by the needs of specific applications.
2. maximum portability: to be able to support various types of implementations, the model must be a conceptual and not a physical model of data. GENELEX pursues exactly the same objectives as the Text Encoding Initiative. The realization of the GENELEX conceptual data model was therefore conceived independently of physical models, permitting GENELEX partners to choose different physical models for the implementation of GENELEX dictionaries.
3. minimum discrimination: there is a need for the model to remain as independent as possible of a particular linguistic theory. For this reason, the linguistic aspect has been reduced to accommodate the needs of all users, who may then, with formal rules of translation and/or deduction, find the information of interest in a GENELEX dictionary.
4. linguistic realism: GENELEX partners attempted to avoid as much as possible to establish constraints on the completeness an electronic lexicon must reach to be declared in conformity with the specification, being conscious that there is still a long way to go before complete and generic lexica are available. The objective of GENELEX is thus a double one:
   - allowing convergence of various existing lexical assets into a common conceptual structure

[3] The partners of this project are ASSTRIL-LADL, Gsi-Erli, IBM France, SEMA Group (France), CONSORZIO LEXICON RICERCHE, SERV. EDI, SOGESS srl (Italy), ILTEC (Portugal), GSI TECSIDEL, UNIVERSIDAD AUTONOMA DE BARCELONA (Spain). The content of this section is Gsi-Erli's responsibility.

- enhancing harmonious developments towards generic and complete lexica by expressing methodological guidelines

The stress put on methodological guidelines leads GENELEX to pay a great attention to explicitation of the meta-linguistic features of the model. There are therefore two main types of components in this model:

- a level for linguistic description, that lists pieces of information ('features') to be attached to lexical entities and defines relations between such entities
- a level for explicit and analytic description of the features themselves, that can be inter-related, structured, organized into hierarchies, refined by the means of more atomic values, etc.

3.3.1.3 Formalisms Nevertheless, GENELEX offers a means of checking the conformity of a given set of lexical data to the model it recommends. The conceptual model is expressed:

- through the entity/relationship formalism, despite of its limits, in order to grant a better communication between partners and to allow a synthetic graphical representation of the model ;
- through an SGML Document Type Definition (DTD) that may serve as a formal specification of the model, and as a format for the interchange of data, which may be checked by an SGML Parser. Large sets of lexical data produced by GENELEX partners have already gone through such controls w.r.t. morphological as well as to syntactical information [4].

3.3.2 Linguistic architecture

The highest level of articulation of the GENELEX model is between three layers:

- morphology
- syntax
- semantics

There is no entity like a "lexical unit" as a whole in GENELEX: any lexical item can be seen either as a progression through the three layers or as a set of information regarding one given layer. Apart from tradition and intuition, the reasons for such a choice lay in what could be called the life cycle of a computational lexicon. A survey of lexical assets of GENELEX partners showed that every other database had reached a different step of development, and that stress had often been put on one of these layers, according to the needs for particular applications and/or to the human resources available. The organization was therefore defined so that it could allow the construction of a GENELEX lexicon to be undertaken starting from any layer (i.e. so that it could welcome data from the partners, after they have been checked

[4] Unfortunately, such controls cannot check the consistency and relevance of linguistic information itself: they merely guarantee that lexical data can fit in a GENELEX model.

with respect to the model of the relevant layer). This, even if the path from morphology to semantics through syntax may be seen as the most likely one. As a consequence, for example, no distinction related to polysemy is formally needed in a GENELEX lexicon untill the semantic layer is being constructed.

A unit of the morphological layer may have zero, one or more related units at the syntactic level ; a unit of the syntactic layer may have zero, one or more related units at the semantic level. A syntactic unit must be related to exactly one morphological unit, except for compound syntactic units, which are not directly related to the morphological units. A semantical unit may be related to one or more syntactic units, (when there are more, all these syntactical units must be themselves related to the same morphological unit). These constraints allow nevertheless a lexical database to focus on the semantic layer and to be more straightforward about the other ones, still being fully compatible with the GENELEX specifications.

In the next sections, a very short description of the type of information that may be found in the three layers is given. A graphical representation in the entity-relationship notation of GENELEX model is proposed in Annex number 5, under a rather general and simplified point of view.

3.3.3 Morphological layer

The morphological layer of the model caters for orthographical and phonological information (as a structural symmetry between Written System and Phonetic System has been found). Its major component is the MORPHOLOGICAL UNIT (UM). Relevant information about written and phonic forms (including abbreviated forms), inflectional behaviour, derivation, internal composition and etymology, may be expressed within this layer. Criteria are proposed for distinguishing several UMs for not-so-straightforward situations.

3.3.3.1 Typology of UMs:

- AUTONOMOUS MORPHOLOGICAL UNITS
  - Simple Morphological Units
  - Agglutinated Morphological Units
  - Compound Morphological Units (frozen compound words): criteria are suggested for those compounds to consider as frozen (vs compound units that could be described at the syntactic level)
- NON-AUTONOMOUS MORPHOLOGICAL UNITS (mainly affixes, but also units that can be found only in compound words)

3.3.3.2 Characteristics of UMs:

- WRITTEN AND PHONETIC FORMS (UMGs and UMPs), with possible mention of their radical(s) ; variants can be expressed through this feature

- GRAMMATICAL CATEGORIES (a list of parts-of-the-speech is defined, with indications about determiners, noun-adjective distinction, status of past participles)
- MORPHOLOGICAL FEATURES, i.e. Combinations of Morphological Features (CombTM) that can characterize a paradigm of inflected forms (Gender, Number, Mood, Person, Possessor Person)
- INFLECTED FORMS
  - Inflectional behaviour of Simple Words (MFG and MFP), that may be described through alternative methods of computation (CFFG and CFFP):
    - addition of an affix to a radical (which reduces the total number of CFFGs and CFFPs)
    - removal or addition of characters for a written or phonemic morphological unit (which relieves the necessity for a morphemic description of UMGs and UMPs)
  - System of inflection of Compounds (MFC), with respect to the behaviour of each component, both internal to the compound and submitted to contextual variation
- DERIVATION is expressed through an ordered set of relations oriented from the derived to its components, qualified according to status (basis, suffix, etc.)
- ABBRIDGED FORMS are expressed through relations between UMs, that may be typed according to the nature of the abbreviating mechanism (acronym, initials, etc.)
- USAGE VALUES, i.e. Combinations of usage values (CombVE) that can characterize UMs (rare, archaic, colloquial, etc.)
- ETYMOLOGY

3.3.4 Syntactic layer

The syntactic layer of the model is composed of SYNTACTIC UNITS (USYN). Relevant information about syntagmatic behaviour of lexical items may be expressed within this layer. This includes:

- information about construction of an item:
  - basic saturated construction in terms of governed "positions"
  - transformational possibilities of such constructions
  - linearity (fixed, restricted, free order of position) when it is not related to stylistic or rhetoric choices
  - pronominalization
  - optionality (mandatory, conditional, optional actualization of positions)
  - distribution of a position:
    - type of syntagm (a limitative list of labels is given)

        * restriction on internal structure of syntagm (such syntagms may be themselves described in terms of positions, so syntactic description may be as refined as desired)
        * restriction in terms of morpho-syntactic features
        * restriction on lexical realization (by the means of aspectual specification, denotational constraints, reference to one particular lexical item, etc.)
        * specification of connecting element
    - grammatical function of the syntagm in a position
    - thematic role of the syntagm in a position

- compound syntactic units, where every component is a 'position', with means for describing all kinds of variations in their surface actualization:
    - the interaction between the "internal" and "external" positions, i.e. how far a syntactic compound is saturated
    - the paradigms of lexical realization of a position in such a compound unit
    - optionality of a position in such a compound unit (mandatory, conditional, optional actualization)
    - accepted transformation(s) (e.g. pronominalization of some components)

The criterion for distinguishing several USYNs related to a single UM is merely the diversity in syntactic behaviours.

3.3.5 Semantic layer

The semantic layer of the model deals with two distinct levels of representation: a first level may be seen as strictly lexical semantics, whereas the second one aims at representing a more cognitive type of contents.

The main entities for lexical semantics are SEMANTIC UNITS (USEM). They are closely connected with the syntactical layer, as every USEM has to be related to at least one USYN. This relation may be restricted through constraints about the USYN itself and/or by filters on the positions it governs in a construction ('productive' collocations can be described by such filters) ; the relation may also precise the way semantic arguments of a predicate match syntactic positions governed by an USYN, and give default semantic values for implicit arguments (for example when an absolute construction is possible).

An USEM may be connected to a linguistic PREDICATE, which can synthetize semantic information about predicative USEMs. This connexion may take various modalities: a PREDICATE may be lexicalized by one or more USEM, so may its arguments. For example, USEM1 cans be one of the lexicalizations of PREDICATE P, USEM2 may be the typical lexicalization of argument i for P, etc.

USEMs can be described:

- analytically through a set of values of semantic features that range from 'classical' componential feature to pragmatic feature referring to a scenario, and include mentions of domain, connotative value, etc.

- by means of cross-references that can be expressed through a set of specialized relations:
  - paradigmatic relations as hypo/hyperonymy, synonymy, meronymy, etc.
  - 'semantic derivation', that may represent derivation according to the meaning (with or without morphologic motivation), or information related to 'typicallity' (e.g. USEM1 is the typical location for USEM2 activity, USEM3 is the typical instrument for USEM4 action, etc.), when not expressed through a PREDICATE
  - relations of collocation preference (e.g. USEM1 is the preferred intensifier for USEM2, USEM3 is the support verb for such aspect of USEM4 action, etc.)

PREDICATEs may also be described in the two manners, although the features and relations available for them are less varied.

The semantic layer should also offer the possibility to abstract cognitive units (CONCEPTs) from USEMs and/or from PREDICATEs. Such units may be useful to factorize information about a lexical class of equivalence (for instance synonyms carrying different connotative values), but also to materialize units of content that do not have a lexical actualization for a given language: such 'lexical gaps' may need to be filled for the purpose of establishing taxonomies of for representing terminological data.

Finally, although there is not within Genelex for the time being a complete specification for a multilingual level of description, the structure of the semantic layer paves the way for:

- a contrastive description of transfer information from lexical semantics in one language to lexical semantic in another language (using 'filters' in the same manner as when relating USYNs to USEMs)
- an interlingua description based upon sharing of CONCEPTs, even if this prospect seems rather fuzzy by now for direct NLP applications, as such a possibility could turn out to be very useful for terminological purposes.

4 Input to the EAGLES architecture work: assessment

4.1 An overview comparison of the Multilex and Genelex models

We give in Annex number 4 and 5 Entity/Relationship representations of Multilex and of Genelex general architecture.

4.1.1 Differences

1. *Information grouping*

   - Multilex has an architecture which distinguishes two nodes (graphical, phonological and morphological information attached to the GPMU, and syntactic, semantic and transfer information attached to the LU node). Genelex has a three-level-architecture, distinguishing a morphological, a syntactic and a semantic level.

- LU (Multilex) corresponds to *USYN* and *USEM* in Genelex.

2. *Variation*
   Form variation is not identically described in Genelex and Multilex:

   - in Multilex, multiple *GPMU*s, identifying orthographic, phonological and morphological sets of attributes, that are related to the same *LU*,
   - *UM* can have mutiple values in each of its level- wise attributes (orthographic, phonological, morphological)

3. *General approach*
   The Multilex architecture is designed as a multilingual language-independent overall structure including all the language-specific models, while the Genelex architecture is abstracted from a French monolingual dictionary model.

4. *Terminology*
   Multilex proposes a model integrating general language and term description.

4.1.2 Similarities

1. *Levels*
   The main entities of both models (i.e. GPMU, LU, UM, USYN, USEM) are seen as nodes of information related to content-wise descriptions or attributes that may therefore be widely shared by many items.

2. *Morphology*
   Both Multilex and Genelex offer means to incorporate an explicit description of the morphological mechanisms that correspond to inflectional models (through the attributes and relations of MFG and MFP in Genelex, through sets of declarations in the *Rule Specification* section of the dictionary in Multilex).

3. *Theory-independence*
   Both Multilex and Genelex can welcome, within the same dictionary, different theoretical approaches and concurrent analyses. The representation language and the descriptive devices do not force the lexicographers to use one approach or another, provided that the pieces of information are translated into the specification. (However, the lexicographer has to control the consistency of his own choices.)

4. *Morphology*
   (Genelex) and *GPMU* (Multilex) cover the same kind of information, i.e., morphology, orthography, phonology;

5. *Compounds*
   (Genelex) and *CGPMU* (Multilex) have the same function;

6. *Pragmatics*
   In both models, 'pragmatic' information (in the sense of usage indication) can be attached to morphology, syntax and semantics, according to the needs

7. *Syntax*
   In both models, *USEM* and *LU* may be related to several syntactic descriptions.

As a provisory conclusion on this quick comparison, it can be stated that:

- On general methodological grounds, as well as in terms of relative role of the architectural proposal, the two projects may be viewed as quite similar.
- Node-wise organization of the lexical descriptions are based in principle on the traditional distinctions of descriptive levels, as they have been established in lexicography.
- There is no separate level of "pragmatics", but pragmatic descriptions are "distributed" over the partial descriptions from other levels.

However, the "packaging" of information is different in the two proposals, which implies that relations between levels partly differ in the two models.
The next sections will discuss these approaches from the point of view of descriptive economy, power of generalization, strategies to accomodate even "hard cases". For the Architecture Task Group, a relevant question to answer will clearly be the status of syntactic information with respect to morphology on the one hand and to semantics on the other hand.

4.2 EAGLES first consensus for a common architecture

The schema in ANNEX 9 summerizes the differences and similarities between the MULTILEX and GENELEX architectures. It should be taken as a basis for further work on the EAGLES architecture.

4.3 Questions to be discussed

The MULTILEX architecture distinguishes two "packages" ofdescriptions of lexical items, GPMU and LU.

The motivations within MULTILEX for having two nodes around which the lexical information is distributed are basically:

- safe extraction of data and (re-)combination of information from/into MULTILEX dictionaries,
- language- and theory-independence of the architecture,
- modularity of each linguistic level linked as a whole entity to one of the nodes which in turn are linked together to produce a full lexical description. Relations between linguistic levels being an exceptional mesure occurring only in "hard cases".
- no need to overload the whole structure to account for exceptional behaviours that can be dealt with using simple appropriate devices.

The main points about the MULTILEX packaging are thus related to the form/meaning distinction and to *maintenance and update.*

The GENELEX architecture distinguishes three "packages" of descriptions of lexical items, UM, USYN and USEM.

The motivations within GENELEX for having the three levels of description are basically:

- gains in descriptive expressivity, in order to welcome lexical data elaborated within different theoretical frameworks,
- focus on syntax as compared with "semantic first" editorial dictionaries, where syntax is often under-specified,
- possibility of a more modular working organization when constructing an electronic lexicon.

4.3.1 Main differences

In this section, the main differences between GENELEX and MULTILEX are analysed in the perspective of a synthesis.
The two models have both adopted the traditional linguistic levels, but they differ in the way they organize them. These differences concern the distribution of information around two or three nodes, and also the cardinality of the relations which rules the functionning of the structure.
Although GPMU and UM seem to be rather similar in their contents and their roles, the cardinality of the links between the node (UM or GPMU) has an important influence on the way the relation between the UM or GPMU on the one hand and syntax on the other hand is dealt with. Namely, while each GPMU identifies sets of unconflicting, homogeneous features and constitutes an autonomous entity, a UM contain variant features needing explicit disambiguating links through the linguistic levels.
The main consequence is the close relationship of UM and syntactic information in GENELEX on the one hand and the autonomy of syntax and GPMU - at least in the great majority of cases - in MULTILEX on the other hand. The difference between these approaches also explains the choice of a three-levels GENELEX or a two-nodes MULTILEX architecture.

4.3.1.1 Syntax In GENELEX, syntactic information is a layer of the model. A node UM can be attached to zero, one or more nodes USYN. A node USYN is attached to exactly one node UM and to zero, one or more nodes USEM, which is roughly represented:

```
UM <--0,n-----1,1--> USYN <--1,n-----1,n--> USEM
```

Syntactic information refers unambiguously to morphology. In MULTILEX, a GPMU can be related to zero (ex: "bacille de Koch" in which "Koch" doesn't need an LU), one, or several LUs. Syntactic behaviours that correspond to nodes LU are described through "syntactic blocks".

```
GPMU <--0,n-----1,n--> LU <--1,n-----1,n--> Syntax
```

Morphology/syntax interaction is therefore described in different manners in MULTILEX and GENELEX. It is necessary to check the possibilities to express the modalities of this interaction within the two models.

4.3.1.2 Accounting for multiwords

1. MULTILEX and SITE/EUROLANG have jointly elaborated the model of the complex GPMU, CGPMU. It contains:
   - the orthographic description of a multiword unit (= a multiword string);
   - the phonetic description of a multiword unit (= a phonetic transcription of the multiword string);
   - pointers to morphological information about those parts of the CGPMU which have themselves GPMUs or CGPMUs in the MULTILEX dictionary;
   - constraints on the elements of the multiword string (no concrete example available, intended to cover e.g. possibilities of modification, insertion, etc.);
   - a relation with an LU.

   This has the following implications:
   - A multiword unit has its own LU and its own CGPMU.
   - At GPMU level, pointers may be used relating the CGPMU with individual GPMUs; so, morphological information need not be spilt out more than once. [5].
2. GENELEX has defined models for describing multiword units both at the morphological level and at the syntactic level. Guidelines for selecting the appropriate level where a multiword item is to be coded are provided.

4.3.1.3 Conceptual description and terminology EAGLES has still not started the work group on terminology. But it is recognized that the architecture should pay attention to the topic of terminology. See minutes of meeting on 16/4/93.

MULTILEX is experimenting with the introduction of "specialized LUs (SLUs)", for the conceptual level. An SLU would be related with several LUs; the LUs related with the same SLU would be quasi-synonymous, and SLU would be a domain-dependent generalization over the meanings of the LUs. SLUs can be cross-referenced.

The same device has been proposed in GENELEX, under the name of CONCEPT.

This descriptive device may be considered as a starting point for work on terminological modeling.

[5]The work on CGPMU was initiated in a MULTILEX/EUROLANG group. The core MULTILEX standard does not fully specify complex word representation. This work could be carried on within EAGLES.

4.3.2 Other issues

A detailed version of guidelines can hardly be produced before the syntax group has discussed the status of syntax information, which has to be decided for an architecture proposal, i.e.

- The main entities of the model
- The articulation between the main entities

The architecture group will need to continue its activity as an infrastructural service for the other groups over the whole period of the EAGLES exercise, to ensure that these recommendations remain consistent with those of the other task groups, and to amend them if necessary.

There should be more work, in both the morphosyntax group and the architecture group, on problems of the representation of multiword units; the following two reasons need to be kept track of:

- even though CGPMU may provide a straight forward way of representing certain types of multiword units, it is not sure that this device is sufficient to cater for the wide variety of phenomena likely to occur. It would be useful to set up a list of the facts about multiword items which need to be treated.
- EAGLES should not commit too early to a given descriptive device for multiword units: it may be better to keep the guidelines flexible and to be able to insert, later, some proposal which may be based on more and also more detailed analysis of phenomena from different languages.

4.3.2.1 Hierarchies, Classification, Inheritance in the lexicon The architecture group considers the introduction of classifcations and, along with these of inheritance hierarchies a useful structuring device on top of flat (and maybe redundant) dictionaries. However, it is less evident whether EAGLES should "prescribe" the use of this device or even of a given instance of it.

Proposal:

- EAGLES recommendations ideally come in two layers:
  - a flat layer of basic lexical descriptions, which may be somewhat redundant, because it does not make use of any classificatory devices;
  - a structuring layer, which introduces classifications, generalizations, etc., and which may involve certain representational ways of expressing these generalizations, e.g. through inheritance hierarchies.
- If an EAGLES user introduces classifcations at the second level, this should be done via specification of relations or generalization rules which express a two-way correspondence between the flat representations and the hierarchical ones. See the discussion about morphological dictionary (base level) and corpus analysis tagsets (applied, second level) in the minutes of the morphology group.

It is assumed that

- different descriptive levels may make use of classificatory devices with different properties (e.g. monotonic vs. non- monotonic inheritance hierarchies);
- different applications of EAGLES lexicons may later introduce generalizations of their own.

If we think in terms of larger actual dictionaries, it would evidently be useful[6] to have some generalization and classification already at the level of basic lexical descriptions, e.g. to avoid redundancy in the different word forms of a lemma (have lemma-specific information instead, and additional ideosyncratic information on word forms separately). The question there is which classifications may usefully be proposed by EAGLES: evidently, there is no single best generalization over a heterogenous set of objects, which is why a generalization process on a "data-only" basis is not possible. Usually, generalizations at this level are based on the classificatory assumptions of the approach advocated.

[6]This topic needs further elaboration, in EAGLES.