EAGLES

Expert Advisory Group on
Language Engineering Standards

EAGLES

Computational Lexicons Methodology Task

EAGLES Document EAG–CLWG–METHOD/B

Version of Jul, 1993

Contents

1 Authors

Ulrich Heid	
Universität Stuttgart	
Azenbergsstrasse 12	Tel: +49.711.121.13.73
HD-70174 Stuttgart	Fax: +49.711.121.13.66
Germany	E-mail: uli@ims.uni-stuttgart.de

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 Introduction

This task is meant to provide methodological guidelines for building the consensus. Its role is to define the objectives and methods for the other tasks and to specify the means of communicating the results of the Working Group inside each task and to the outside.

An important aim of this task is to define the scope of the EAGLES standardization entreprise for the Lexicon Working Group, its limits in width and depth, viz.:

- exhaustive description vs. basic consensus features
- all languages vs. a restricted list of languages
- empirical observations vs. application/theory-oriented approaches
- redundancy vs. non-redundancy
- etc.

The answers to these questions determine the overall orientation of the different tasks and the basis for an evaluation framework. In order to perform the necessary tasks of circulation and integratability checking of the work done at each linguistic level, the methodology group should be active over the whole period of the EAGLES exercice. This report consists of three parts: Objectives, basis for comparison and presentation of the results.

3 Objectives

- The EAGLES descriptive standards proposal is made for the construction of a lexical resource, thought of as being multifunctional, but not for a given application resource or for a given range of applications. This has to be kept in mind in discussions about generalization devices, about the use of hierarchical or other types of structures superposed onto the EAGLES descriptions, questions of redundancy or elimination of redundancy, and so on. This is also related with questions of the formal representation of proposals which will be the output of EAGLES work.

- The experience acquired in the different projects dealing with lexical knowledge modelling has shown that it is not realistic to expect the same degree of consensus for all levels of linguistic description. Typically, while arguments about orthography and morphology rest on well established linguistic facts, arguments on semantics rest on theories for which no experimentation has been extensive enough to make sure that the descriptive devices used are already defined on the basis of sufficient uncontroversial facts. For this reason, the methodology applied to each level should be adjusted to the kind of results that it is likely to produce.
  The levels of linguistic description for which it is most realistic to come up with proposals for descriptive guidelines in the short or medium term should be considered with more emphasis than the other ones; morphology and morphosyntax are among these levels, see the revised workplan for that matter.

- It is not relevant to determine priorities on languages to define standards for. The project will concentrate on overall features for the languages of the EEC, while the possibility of representing other languages – Swedish, for example – should not be excluded.
  A comprehensive description of lexical features for all european languages is a long task which would entail the participation of a specialist of each language in each task dealing with each linguistic level. The EAGLES objective is neither to provide such a comprehensive description nor to concentrate on specific languages. Such language-oriented studies may lead to unsatisfactory results such as, for example, a well developed morphology in German and French and very little in English and Spanish while syntax would be completed only in French and Italian according to the expertize available in each Task Group.

- A core-extensions approach should be applied in order to quickly bring out a first consensus and prepare the next steps. Ideally the construction of proposals for guidelines for individual linguistic levels should be spiral-wise, in the sense of making available a first draft as soon as possible, getting feedback, amending and improving the proposal, circulating it again, etc.

  For the other levels, the work of the EAGLES lexicon group should lead to a survey of the available approaches, as well as to a methodology of comparing these proposals. Making such a methodology available and testing it on a relevant subset of phenomena and approaches is a goal in itself. A starting point may be work in the EUROTRA-7 Study and in the NERC project.

The goals should be discussed by the participants to the task groups and quickly decided for at the very beginning of the task work. Here follows a proposal for goals for each task. This proposal is a general framework within which each task will fix its own limits according to the input material provided by a survey of available proposals.

- Architecture: core linguistic model, definition of the lexical levels and their relations.

- Orthography: core of consensus features and comparison of specific extensions (to be clarified later),

- Phonology: information needed by NLP applications,

- Morphosyntax: core of consensus features and comparison of specific extensions,
- Syntax: core of consensus features and comparison of specific extensions (to be clarified later),
- Semantics: aspects identified where steps towards consensus building can most realistically be taken; directions for future work,
- Multilingualism: state of the art connected to the results of the other levels (to be clarified later),
- Terminology: state of the art connected to the results of the other levels (to be clarified later).

4 Basis for comparison

A major task of methodology is to determine how to efficiently represent information related to the objects dealt with. In this regard, description and representation devices of three different levels have been identified for EAGLES.

4.1 A support language for conceptual modelling

In order to exchange questions and comments between the EAGLES partners and to evaluate progress, the entity/relationship model should be used as a means of representing the organisation of lexical data. Both the MULTILEX and GENELEX linguistic architectures allow for entity/relationship representation. However, this should not be interpreted as a decision of the WG to restrict the expressive power, in the modeling of its lexical descriptions, to that of a particular storage and representation device, such as relational databases (which are usually described in entity/relationship terms). The entity/relationship representation is used because of its transparency and wide use.

4.2 An interchange format

A format is needed to exchange electronic versions of lexical descriptions in a way that preserves their structure and allows to check their consistency accross platforms and representation systems. This format should be based on SGML (Standard Generalized Mark-up Language). The main advantages of SGML are well known:

- SGML is an ISO standard,
- SGML allows to define and check a generic structure for any kind of document,
- The ASCII files it defines are portable in all environments.

4.3 A linguistic formalism

The EAGLES working group on formalisms should be invited to collaborate with the subgroups of the WG (for example, the WG on Lexicon Architecture) to decide whether a full representation of EAGLES lexicon-group results in a linguistic formalism is possible and needed and which of the candidate formalisms – TFL or another one – is suitable for the representation of lexical data. Some members of the WG think that this types of formalism is more NLP application oriented or theory oriented and should be derived from the LDB by an export process. The interaction between the three types of representation languages is schematically represented in the illustration in figure (Fig. 1).

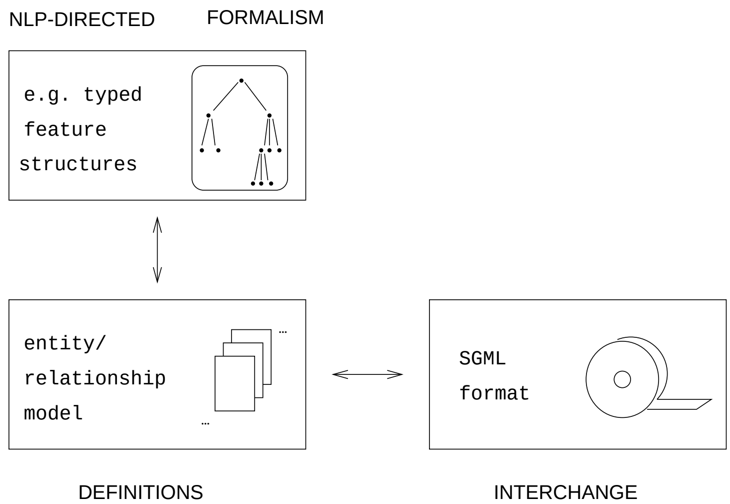

Figure 1: The three types of representations used for EAGLES lexicon descriptions and their interaction

4.4 intertranslatability

The major problem of the interaction between these three formats is that they differ in expressive power, formal properties and with respect to the existence or non-existence of an interpretation in certain types of models.

In other words, intertranslatability is not a priori given, for the three types of description languages. The Genelex project use an ER model to represent the conceptual model. This formalism express a lot of integrity constraints via the typing of objects and relations and through the arities of the relationships. To maintain the model readeable the constraints on cooccurence of features values are expressed in a pseudo-natural language rules. All these constraints are explicitely reported when translating the EA schema in the SGML DTD. This translation follows strict rules. The methodology group will interact with the EAGLES formalisms group and with entity relationship model experts in the consortium to precise the interrelationships between the three representation formats, becausethe level-wise workgroups will want to represent their proposals in a consistent way which then can be interpreted consistently across groups, and also from outside EAGLES. **This point needs further clarification**.

5 Presentation of the results

For each task, results should be presented in three stages.

5.1 Survey of available proposals/standards

The following actions will be performed in the survey phase: the results available so far from projects aiming at proposing guidelines or standards for lexical description and representation will be put together into a survey document. The languages covered include all languages of the EEC, with the possibility of having extensions, for example towards Swedish. If for a given level descriptive devices are proposed which apply only to a given language or set of languages, these languages are indicated.

For the levels of functional syntax and semantics, the outcome of the survey should be a list of the linguistic facts treated by the standards-oriented projects and within existing large lexicons, as well as an assesment of the possibilities of reaching a consensus over the description of these facts.

5.2 Comparison

Methodology for comparison should be regarded as an achievement in itself because it raises difficult questions:

- How to decide on adequate criteria for comparison?
- Comparison for what purpose, for what kind of evaluation?
- How to extract from the specific approach of each proposal a set of comparable features?

5.2.1 Description other project approaches to comparison

- In EUROTRA-7, it was proposed to relate each complex descriptive device with a set of operational tests which can be performed on corpus material in order to make sure that the comparison can be carried out at the level of facts observable in corpora.
- In the NERC Study, a preliminary comparison of morphosyntactic description schemes in corpus tagging has been carried out. The underlying principle in view of a comparative assessment of the facts covered by the individual tagging schemes is to allow for different levels of granularity in the description and in the respective tagging. For example, part-of-speech tagging may be refined with morphosyntactic annotations, and one approach to tagging might restrict itself to a subset of the inventory of morphosyntactic features which are morphologically marked in a given language, whereas another one might try to be exhaustive with respect to morphologically marked and thus corpus-observable differences. The fact that morphosyntax for a given language is mostly a problem of different classifications of a finite domain allows for the comparison of classification schemes (via their tagging schemes) by use of the variable-granularity approach.

  A different methodology has been chosen in NERC for the syntactic area. There, the problem is to be able to relate complex syntactic descriptions, for example from the functional level, with their corpus-observable (and thus intersubjectively reproducible) categorial and morphosyntactic reflexions. Part of this work will be carried out in the DELIS project, where a number of exemplary verb classes will be syntactically and semantically described and descriptions will be related with corpus-searchable evidence.

5.2.2 Consensus and recommendations

- It seems that individual methods of comparison and different matrices for comparing the phenomena covered and devices used by the approaches observed are necessary for the different levels of linguistic description. Once these comparison criteria have been established, it is possible to determine commonalities and differences between the approaches. Using the comparison grid, the next step must be to point to the set of shared basic descriptive devices and to the set of core phenomena catered for in the different approaches; in a similar way, the comparison must point out those areas of description, where either there is no possibility to relate different approaches, or only a few approaches are at all available which take the respective phenomenon into account. The areas where commonalities are observed are those which lend themselves to the formulation of proposals for consensual guidelines.
- In the areas, where no consensus can be achieved, at least guidelines for choosing among alternative, although mutually exclusive approaches should be given, such as to give the user a possibility of choosing on a basis of a definition of his individual needs. This latter point will be performed on the basis of the experience of the participants, since it is difficult to give operational criteria.

The production of recommendations is thus seen as a successive filtering process, and the survey phase is a first input to the comparison of alternative approaches. The survey itself can evidently make use of previous surveys carried out in different frameworks.