Standards to Make Natural Language Resources Shareable Resources

Nicoletta Calzolari Antonio Zampolli
Istituto di Linguistica Computazionale del CNR, Pisa
Dipartimento di Linguistica, Universita' di Pisa
Via della Faggiola, 32
56100 Pisa, Italy
Tel: +39-50-560481
Fax: +39-50-589055
E-mail: glottolo@icnucevm.cnuce.cnr.it

Abstract

Natural Language Resources which are effectively shareable, and possibly usable by the same basic tools, should be designed and produced according to a common set of consensual and agreed specifications, i.e. standards. We can consider the provision of standards as a sort of "metaresource" itself. In this paper we present a broad overview of the European efforts towards standards, describing the LRE project EAGLES.

1 Why contribute standards together with Natural Language Resources?

If we want to produce and exchange Natural Language (NL) resources which are effectively shareable, and possibly usable by the same set of basic tools, with results of processing which are really comparable and integrable, we should aim at designing and producing these resources according to a common set of consensual and agreed specifications, i.e. standards.

In this paper we present a broad overview of the European efforts towards standards, and concrete details on the first results - after one year - of a project launched by the CEC DGXIII in the framework of the LRE (Linguistic Research and Engineering) Programme, EAGLES (Expert Advisory Group on Language Engineering Standards).[1]

This initiative is the result of extensive consultation with leading industrial and academic centres, and of recommendations made by the CEC's Language Engineering strategy committees.

Its objective is to accelerate the provision of standards for the development, exploitation and evaluation of large-scale language resources, such as text and speech corpora, grammars and lexicons, together with computational linguistic formalisms for expressing and manipulating linguistic knowledge and evaluation and assessment methods.

[1] EAGLES consists of five Working Groups hosted by designated R&D centres (Instituto Cervantes - Spain, GSI-ERLI - France, DFKI - Germany, CST - Denmark, Vocalis Ltd. - U.K.) and coordinated by the Consorzio Pisa Ricerche, Italy.

We could consider the provision of standards as a sort of "metaresource" which, on the one hand, is being designed and constructed by a common effort of the major actors in the field working in Europe and, on the other hand, is immediately made available to the interested communities to be tested, validated, revised on the basis of feedback, etc.

2 The EAGLES initiative

The flowering of so many projects on large linguistic resources, first in the lexical area already in the late 1980's, then in the field of text corpora, made the linguistic communities aware of the necessity of avoiding costly and ineffective reiterate constructions of e.g. system-dependent lexicons.

Reusability of linguistic resources has therefore become one of the leading paradigms in the R&D community. Reusability of a linguistic resource, however, presupposes agreement on a *neutral* shared linguistic model and a language to represent such a model. Developing common specifications for linguistic phenomena and for linguistic representation is therefore crucial in order to i) pave the way towards the construction of commonly usable basic linguistic resources, ii) enable merging of material developed in different sites and coming from different sources, and iii) create the conditions for developing common and publicly available tools based on the common model.

3 Setting up of EAGLES

The situation sketched above created the appropriate conditions for the launching, firstly, of a project definition study in 1992 (EEG, i.e. European Expert Group) and then, in February 1993, of a project aiming specifically at defining standards or preparing the groundwork for future standard provision, i.e. EAGLES.

The areas of concern to EAGLES are text corpora, computational lexicons, grammar formalisms, evaluation and assessment, and spoken language. In each of these five Working Groups (WG) leading experts of both the research and the industrial communities are represented, conjoining their efforts towards the development of a common basic European infrastructure and of agreed linguistic specifications. All the major European projects on linguistic resources are represented in EAGLES.

Relevant common practices or upcoming standards are being used where appropriate, particularly in the areas of lexicons, text encoding, and speech. Numerous theories are being taken into account, where appropriate, as any recommendation for harmonisation must take into account the needs and nature of the different major contemporary theories. EAGLES is also drawing strong inspiration from the results of major projects whose results have contributed to advancing understanding of harmonisation issues.

3.1 General Mode of Operation of EAGLES

The basic idea behind EAGLES work is for the group to act as a catalyst in order to pool together concrete results coming from current major European projects. The major efforts go into the following types of activities, which, in this sequence, show how, on very general lines, the work

is organized in all the five WGs. A schematic account of the general methodology of work is outlined below:

- Surveying and assessing available proposals of shared specifications in order to evaluate the potential for harmonization, convergences, and emerging standards.
- Assessing and discovering areas where there is a consensus about existing linguistic resources.
- Detecting those areas ripe for short-term standardization vs. areas still in need of basic research and development.
- Proposing common specifications for the core set of basic phenomena on which a consensus can be found.
- Setting up guidelines for representation of core sets of basic features.
- Drawing up feasibility studies for less mature areas.
- Suggesting actions to be taken for a stepwise procedure leading to the creation of multilingual reusable resources.

4 Methodology of work and first results in the Lexicon and Corpus Groups

4.1 Methodology of work

We give here some ideas on the methodology of work of two of the EAGLES WGs, the Lexicon and Corpus WGs.

Both WGs are obviously not starting from scratch. There are numerous projects which they have built upon, such as ACQUILEX, ET-7, MULTILEX, GENELEX, NERC, TEI, just to mention a few.

In both groups, in the first phase of Survey and Assessment of what exists for each investigated area (e.g. text representation, text typology, morphosyntax, semantics, linguistic annotation, etc.), issues such as the following are considered:

- existing categories and features, also according to the most prominent linguistic schools;
- linguistic phenomena to be described (taken from linguistic descriptions, existing lexicons, corpus analysis, etc.) according to the ET-7 methodology of reaching the layer of minimal granular distinctions;
- existing practices in NLP;
- formalisms for representation of information;
- requirements of NLP applications.

The assessment phase must lead to an evaluation of:

- the feasibility - and up to which level - of building on and reusing what has already been achieved;
- areas not covered so far or requiring more work;
- the adequacy of what exists vis-a-vis linguistic schools.

The Survey Phase is followed by a Proposal Phase, where all the above parameters, and their interaction, are taken into account in the process of designing a common proposal for a core set of features and common specifications which satisfy the various requirements. In particular:

- For each feature, tests and reproducible criteria for linguistic classification have to be devised.
- An evaluation must be made of the advisability of defining a minimal level of linguistic encoding for a computational lexicon or a text corpus to be considered as reusable and conformant to EAGLES specifications. This level could then be applied for evaluation of future lexicon and corpus building projects.

Moreover, the applicability of the results to all European languages has to be pointed out.

4.2 First results

We sketch out here some preliminary results obtained in some areas, chosen among others as being of more interest to the Workshop.

4.2.1 Lexicon

General principles concerning the Lexicon proposals are the following:

- EAGLES proposals concern *multifunctional* lexical resources, not just a given application. They are meant as high-level guidelines, similar to a schema, not as full-fledged dictionaries;
- given the orientation towards multifunctionality, the elimination of redundancy within an EAGLES lexicon definition is not a priori a major goal; devices for lexicon structuring and organisation may be added on top of EAGLES standards proposals, but need not be included in these;
- different descriptive levels allow for quite different degrees of detail of standardisation;
- a core/extensions approach is followed.

As far as the level of morhosyntactic encoding is concerned, a detailed comparison of GENELEX, its application dictionary Alethdic, MULTILEX, NERC and the Corpus WG proposal has been completed, and a first EAGLES proposal has been drafted accompanied by detailed applications of the European languages. The goal is:

- to show overlaps and differences in the inventories of morphosyntactic attribute/value assignments for word form types;

- to come up with a core-and-extensions-approach for the definition of attribute/value descriptions of wordforms as the starting point of an EAGLES proposal;
- to check applicability of the above proposal to the languages involved in the EAGLES group;
- to come up with the conceptual core of mapping statements relating individual existing morphology systems with the proposed set of attributes/value declarations.

The first EAGLES proposal concerning lexical specifications at the Morphosyntactic level, based on a careful analysis of existing practices consists, for each category/POS, in an enumeration of all the potentially relevant features (attributes and possible values) for a number of different European languages.

Based on a comparison of the results, a first inventory of labels has been proposed by the EAGLES group. The EAGLES proposal is organized into a number of layers (or levels), with increasing granularity and a distinction between (1) descriptive devices used for all European languages and (2) language-specific extensions.

Concretely, the following classification is used:

- Level 0 (L0):
  - information type: part-of speech classification;
  - recommendation type: obligatory;
- Level 1 (L1):
  - information type: morphosyntactic information of the "agreement features" type ("grammatical" information), applicable to *all* languages;
  - recommendation type: recommended; these features are proposed as a "minimal common core set of features" for the linguistic objects of each category;
- Level 2a (L2a):
  - information type: morphosyntactic information applicable to *many* languages (at least 3); refinement of L1 features, where possible and accepted as useful;
  - recommendation type: optional; we assume that the features are consensual, useful, and relatively easy to standardize; they are, however, not yet part of the practice of most projects. For that reason, EAGLES proposes their inclusion into a description, but the proposal cannot be binding.
- Level 2b (L2b):
  - information type: language specific decriptive devices, for (possibly) different levels of description;
  - recommendation type: optional; language groups may decide on the inclusion of such attributes into the language-specific adaption of the *ELM proposal.*

The above proposals have been applied to a large number of different European languages, with the goal of:

- ensuring the completeness of the general (cross-linguistic) part of the EAGLES proposal;
- spelling out in more detail the language-specific part;
- performing a comparison of the EAGLES proposal with the practice of the language groups represented in EAGLES.

4.2.2 Corpus

The overall goal is to work towards providing guidelines for various aspects of corpus construction, representation and processing, with the ultimate aim of ensuring the interchangeability and reusability of the data and the usability of common software tools.

As far as Corpus Typology is concerned, draft definitions of corpus terminology have been proposed for key terms, particularly those surrounding the notion of corpus itself. It has become clear that terminology is not stable, and that the central word *corpus* is used in a range of meanings so broad as to be sometimes incompatible with each other. Hence early attempts to define terms individually have given way to a more principled approach.

For Text Representation and Encoding, the group has laid out the issues to be addressed in order to develop an encoding standard for corpora, providing preliminary ideas on:

- scope of the corpus standardisation work (definition of a corpus, which text types, languages, facts to encode, etc.);
- relation of this work to the TEI (what they provide, how we need to build on or modify TEI results, etc.);
- definition of levels of standardisation (metalanguage level, markup specification level, markup use level);
- criteria for corpus encoding standard design (e.g. completeness, consistency, recoverability, validability, compactness, readability, capturability, processability, extensibility);
- facts to be encoded (levels of encoding, minimal requirements and conformance for each level, etc.).

In the area of morphosyntactic annotation of text corpora, where the work has been carried out in close cooperation with the Lexicon group, the need to use a language-neutral form, or *intermediate category set*, for the representation of categories by means of morphosyntactic tags has been recognized. This category set would be intermediate between the set of categories specified for the lexical database, and the set of tags used in any specific language annotation task. A convenient linear method of representation has been arrived at, and is being validated by subgroup members and correspondent experts for its application to EC languages.

Three degrees of constraint have been proposed for the description of word categories by means of morphosyntactic tags:

1. *Obligatory features*,
2. *Optional features*,
3. *Special extensions*.

5 Validation and dissemination

By its nature, EAGLES must interact closely with the scientific and industrial R&D community not only in Europe, but world-wide. Since EAGLES involves many bodies active in European NLP and speech projects, close collaboration with these projects is assured and, in many cases, manpower has been contributed by the projects. Procedures have been established allowing EAGLES to access relevant material developed by EAGLES participants working in other projects. Moreover, close collaboration exists with relevant LRE and other European projects not directly represented within EAGLES.

Comments and reactions to the reports produced in the first phase will be fed into the work of the second phase.

Validation of the results will be confirmed by the extent to which they are adopted by the community. Given the initial composition of EAGLES and its feedback mechanism, to be employed at the end of the first phase, it is expected that adoption of EAGLES recommendations will be rapid and substantial. The morphosyntactic proposals are already being applied and tested in a number of LRE projects.

6 Some concrete information

A set of EAGLES Reports will come out in the autumn and will be distributed to selected projects and bodies. Their availability, modes of distribution (ftp, EAGLES server, etc.), status of development, etc. will be detailed at the Workshop.

For any inquiries contact Tarina Ayazi at: `eagles@icnucevm.cnuce.cnr.it`.

References

[1] N. Calzolari, J. McNaught (eds.): "EAGLES Second Progress Report". Internal report, November 1993.

[2] U. Heid, J. McNaught (eds.): "Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications". Eurotra-7 Final Report, Stuttgart, 1991.

[3] G. Leech: "Draft partial report on task 3: Assessment of of recommendations for linguistic annotation. Subtask 3.1 Assessment of recommendations for Morphosyntax level". Draft technical report, Lancaster, December 1993. EAG-CSG-T3.5.

[4] Lexicon Working Group: "WG Computational Lexicon Workplan". Internal Report, Pisa, 1993. EAG-LWG-WP.

[5] M. Monachini, N. Calzolari: "Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora and Applications to European Languages". Draft Technical Report, ILC Pisa, January, 1994. EAG-LSG-T4.6/CSG-T3.2.