# INFORMATICA UMANISTICA
# DALLA RICERCA ALL'INSEGNAMENTO

**Atti Convegni *Computer, Literature and Philology***

**Roma 1999 – Alicante 2000**



Universitat d'Alacant
Universidad de Alicante

a cura di Domenico Fiormonte

(con la collaborazione di Giulia Buccini)

ESTRATTO

BULZONI EDITORE

# STANDARDS FOR LANGUAGE DATA PROCESSING: AN HISTORICAL OVERVIEW

*Antonio Zampolli*

## 1. The context of standardisation efforts: the historical evolution of the field

The awareness of the needs for standards in the field of electronic language data processing (LDP) goes back to early times, and the search for some form of standard must be considered in the context of its historical development.

### *1.1 The two initial sub-fields of LDP*

It is well known that the field LDP started a few years after the end of World War II, with two major sub-fields:

- Machine translation (MT), in particular for military applications (Booth *et al.*, 1958);
- Electronic lexical text processing: (indexes and concordances, Busa 1951) for various humanistic disciplines (philosophy, literature, etc.): Humanistic Text Processing (HTP).

In the first 15 years, the two sub-fields were aware of their complementarity, of various common scientific and technical problems, and contacts and co-operative exchanges were frequent and fruitful.

Around the mid-60's, the contacts between the two sub-fields became less and less frequent.

As a consequence of the ALPAC Report 1966, MT projects lost the support of the Funding Agencies, and almost all of them were closed – with very few exceptions – both in the USA and in Europe.

The ALPAC Report explicitly stated that MT should be replaced by a new discipline, for which the suggested name was: "Computational Linguistics" (CL).

### *1.2 The divergence between HTP and CC*

Despite the recommendations of the ALPAC Report, which stated that CL should include among its priorities assembling and studying large corpora, lexica, and grammars at the monolingual and contrastive level, CL started to focus almost exclusively on models of a few monolingual syntactical phenomena. Rather than on building large corpora, lexica, grammars, robust procedures for processing 'real' language uses in 'real' communicative contexts, the focus was on studying some formal properties of specific linguistic theories or some algorithmic mechanisms, and on testing them with a few 'critical' data, selected ad hoc for their relevance to the task at hand or with a few artificially constructed examples (Godfrey / Zampolli 1977: 382).

On the other hand, HTP was dealing with an increasingly large quantity of texts, and the major efforts were directed to exploiting the benefits of new developments in hardware technology. But the processing was usually limited to the level of graphemic units, without taking advantage of knowledge and methods, developed in CL, for identifying and processing units recognised at various levels of analysis (lemmas, morphosyntactic features, etc.).

Only a few Centres, in compliance with their institutional mandate and inspired by the vision of some researchers, were active in both sub-fields, successfully pursuing methodological and practical cross-fertilisation.

### *1.3 A new co-operation paradigm*

In the second half of the 80's, the recognition and spread of the so-called 'language industries' (LI: the term was consecrated at the Conference of the same name organised in Tours in 1986 by the Council of Europe) (Vidal-Beneyto 1991) created the conditions for a new mutual interest and a concrete co-operation between CL and HTP.

In this context, HTP could offer CL its know-how and heavily tested tools for the design and composition of large corpora, selection and encoding of texts and textual features, identification

and study of styles and sub-languages, lexicographic documentation and analysis, quantitative research; whereas CL could offer HTP know-how and tools for morphological analysis (adaptable to diachronic linguistic changes), morphosyntactic taggers (statistical or rule-based), shallow parsers, automatic (parallel) text alignment, learning and 'discovery' procedures for extracting linguistic knowledge from corpora, 'intelligent' linguistic text browsers for information retrieval/extraction, lexical/conceptual/knowledge base, which can be used, e.g., for lemmatisation or for expanding the 'recall power' of human queries in textual browsing (cfr. Calzolari / Picchi 1985).

### *1.4 Language Resources*

The term 'Language Engineering' is increasingly being used: it underlies the engineering efforts (coverage, robustness, and adequacy to the real data to be processed by the applications) needed to move from prototypes to real-life products.

An essential pre-condition is the availability of adequate Language Resources. This term (which I introduced in 1991 in a Panel of EC appointed experts in charge of designing a strategy for future developments, and which is now widely used in the literature and in the Funding Agencies' programmes) usually represents large collections of language data and descriptions in Machine Readable Form (MRF) used for building, improving, evaluating algorithms and systems for natural language (spoken and written) processing: for example, language corpora, documenting real language usage, and computational lexicons, providing knowledge about the linguistic properties of hundreds of thousands of lexical entries.

The turning point was the Grosseto Workshop (1986) (Walker *et al.* 1995), where the research community recognised the central role of LR in the development of CL, the complementarity of the rule-based and the data-based approaches in NLP, the need to replace the study of a few critical data with the study of the variety of linguistic phenomena occurring in real communication.

Among the final recommendations (Walker *et al.* 1987), it seems appropriate to quote two in particular: i) exploring the feasibility of creating multifunctional lexical databases capable of general use, despite the fact that CL systems can use different linguistic theories and different computational and applicational frameworks; ii) studying the possibility of linking lexical databases and large text files, in both monolingual and multilingual contexts.

### *1.5 The need for consensual, de facto standards*

The costs and efforts involved in creating adequate LR demands for:

(i) strong co-operation between the various researchers and organisations interested in the availability of LR in both the sub-fields, HTP and CL. In fact, ALLC, ACH, ACL have begun organising co-operative initiatives in recent years;

(ii) sharing, at the organisational level, the cost and effort between different researchers (lexicographers, linguists, corpus linguists, computational linguists, LI developers, psycho-linguists, etc.), and a variety of organisations: public research Institutes, Universities, private Companies, national Governments, national and international Funding Agencies, etc.

The Pisa Group, for example, has promoted and is currently co-ordinating (i) two projects of national interest for the creation of a national infrastructure of LR for Italian, with more than 30 partners, public and private; (ii) the PAROLE/SIMPLE EC projects aiming at building a set of comparable corpora and harmonised computational lexica for 12 EC languages; (iii) the co-ordination group (ENABLER) of European national projects for LR, with the participation of 14 Institutes of different European countries;

(iii) promoting the sharing and widespread availability of LR. We founded ELRA (European Language Resources Association) aiming at a large diffusion of available LR. ELRA will actively

co-operate with LDC (Linguistic Data Consortium), the equivalent organisation in the US;

(iv) ensuring the reusability of LR for different goals (multifunctionality) in different applicative and theoretical frameworks (polytheoreticity). To this end we have promoted the adoption of common specifications, produced on the basis of a consensus among the major elements in the field, proposing and co-ordinating the EAGLES standardisation project: Expert Advisory Group on Language Engineering Standards.

## 2. TEI and EAGLES

This brings me to the central topic of my paper, namely a brief summary of the historical development of standardisation efforts.

My purpose is (i) to describe the background and development, and (ii) to compare the goals of the two major current standardisation initiatives: the TEI (Text Encoding Initiative) – focusing on HTP, and EAGLES – focusing on CL.

### *2.1 A brief history of standardisation efforts in LDP*

The interest for standardisation dates back to the early times of LDP. In this period (about 40 years), the focus shifted, following the evolution of computing and LDP technology.

In 1961, the participants to the 'Colloque International sur la Méchanisation des Recherches Lexicologiques' (organised by Bernard Quemada in Besançon) (Quemada 1961) discussed the need propose to the hardware producers the adoption of a common enriched set of characters for encoding texts on the mechanographical support.

The problems of representing the variety of graphemes to be encoded were also discussed at a workshop organised by IBM in 1964 (Kay 1964), and Martin Kay published an article (in "Computers and the Humanities") on "Standards for encoding data in Natural Language", taking into account the outcome of this workshop (Kay 1967).

Starting in 1965, practically all the numerous Italian Institutes aiming at using electronic methods for the lexical analysis of texts (cfr. Zampolli 1973, for a list of these projects) began asking the Divisione Linguistica del CNUCE to provide the software for their projects. I quickly understood that, in order to avoid writing many different software packages to accomplish similar operations on texts, I had to design and prepare a set of basic software components, which could be combined in various sequences to form different procedures, thus satisfying the needs of different types of users. The feasibility of this approach obviously required as a precondition the use of a common encoding scheme, capable of representing different types of texts for different languages in different alphabets, and guidelines, ensuring the harmonised application of this de facto 'encoding standard' (Bindi / Orsolini / Zampolli 1979).

As the practice of performing lemmatisation, in particular for comparing at qualitative and/or quantitative levels different texts or corpora, widespread among Italian researchers, I tried to enrich the 'standardised' text encoding scheme with common criteria to be adopted in lemmatising Italian and Latin texts.

Starting with the observation that differences in lemmatisation practices mainly consisted in the different granularity of the distinctive features adopted for identifying and distinguishing a lemma and its form, I proposed an hierarchical organisation of the features used (which can be exhaustively enumerated), in which each node corresponded to a level of granularity for a given feature. The system allowed the user to automatically establish correspondences (via inclusion relations) between lexical units identified, by different projects, at different levels of granularity.

The DMI (Italian Machine Dictionary), built in these years (begun in 1968) according to this system, could thus be customised to perform the lemmatisation following the specific linguistic/lexicographic criteria of different projects, ensuring at the same time the comparability of their results (Zampolli 1976).

A feasibility study to explore possible ways of ensuring the exchangeability of texts across groups of research centres in different countries (Italy, France, Belgium), was promoted by myself, Bernard Quemada (Besançon) and Paul Tombeur (CETEDOC, Loeven) in the second half of the 70s.

A major step towards the promotion of an international standard for text encoding was achieved through the recommendations, proposed (by myself and B. Quemada: Quemada / Zampolli 1981) and approved at the final session of the Workshop on 'Possibilities and limits of computers in producing and publishing dictionaries', organized in Pisa in 1981 at the request of the ESF (European Science Foundation). This session was attended by representatives of the ESF Standing Committee on the Humanities, the Research Funding Agencies of the countries affiliated to the ESF, and the US NEH (National Endowment for the Humanities) (Zampolli / Cappelli 1983).

The need for standardisation of text encoding was thus presented to the Funding Agencies of the major countries, which became increasingly aware of the relevance of this issue.

An even more decisive step was that of the conclusion of the above-mentioned Workshop on 'On automating the lexicon', held in 1986 in Grosseto, near Pisa, promoted by the committee of experts of the CEE for NLP (CETIL), the University of Pisa and the Institute of Computational Linguistics of the National Research Council, and sponsored by the Council of Europe, ESF, ALLC, ACH, ACL, AILA, EURALEX, etc.

The papers and the discussions presented at the Workshop made clear that:

- the lexicon should be considered a central component of NLP systems and linguistic models;
- lexical information is vital for different disciplines and research: linguistics, CL, HTP, AI, anthropology, psycholinguistics, ethnology, literary research, cultural studies and history, etc.;

- the average size of the then available computational lexicon was 12 lexical entries (sic!). These lexicons were examples intended to show structural and definitional problems, and were not really usable resources. Furthermore, the few existing lexica of a realistic size were highly 'idiosyncratic', so that the same researchers, even within the same company, were forced to start from scratch to build a totally new lexicon for any new application they were aiming at: the cost and the effort of this duplication was difficult to accept;
- at first, it seemed it would have been to neutralise the varieties in lexical information provided by different lexical models, largely due to idiosyncratic 'stylistic choice of formal devices of different linguistic schools', without losing the lexical information content;
- computational lexica should be seen as components of a basic structure, also including large textual corpora, associated tools, and the necessary organisational facilities.

In conclusion, the motivations for standardisation evolved following the evolution of the overall technical and organisational context of LDP:

- problems of the physical representation of characters in mechanographical encoding (1960-65);
- the need to reuse common software tools (at an early stage, it was difficult to count on programmers for humanities and linguistics computing) (1965-70);
- the need for a reusable/shareable linguistic analysis of lexical data (1970-80);
- increasing awareness of the need for a textual encoding standard for text exchange, and a large HTP research community (ESF – 1981);
- need of reusable lexica and corpora to ensure the 'robustness' of the applications required by language industries (Tours – Grosseto, 1986).

## *2.2 The current Initiatives*

### 2.2.1 EAGLES

In order not to lose momentum, the morning after the end of the Grosseto Workshop I called a 'working breakfast' of young researchers attending the Workshop, asking them to analyse and discuss the possibility of designing guidelines/standards for a reusable polytheoretical lexicon.

This group became the nucleus of the so-called 'Lexicon Pisa Group', which worked from 1986 to 1988, with the participation of representatives of certain major linguistic schools, whose theories and models were then inspiring NLP system builders: GB, GPSG, LFG, RG, SG, etc.

Part of the work of the Lexicon Pisa Group had an experimental nature: a few lexical entries, mainly verbs, were selected; each school representative presented his description of the various syntactic constructions of each entry; the descriptions of the same lexical construction were compared and the group tried to formulate a commonly agreed 'neutral' description, from which all the descriptions proposed by the different schools were derivable/computable by means of an automatic conversion process (Walker *et al.* 1987).

These experiments were successful: encouraged by the positive results, I had the opportunity to propose to the CG12 a project to assess the feasibility of a 'polytheoretical' lexicon at various linguistic levels. The CEE DGXIII accepted the proposal, launching the project known as ET-7, co-ordinated by Ulrich Heid of the University of Stuttgart.

In the meantime, various projects were launched, in different European research frameworks, in order to respond to the needs for computational lexicons of the R&D community:

ACQUILEX (ESPRIT – Basic Research, co-ordinated by A. Zampolli at the University of Pisa) aimed at exploring the possibility of (semi)automatically extracting lexical information relevant for NLP applications from machine-readable

dictionaries, in particular, semantic taxonomies and other semantic relations from the parsing of the definitions.

GENELEX (EUREKA, co-ordinated by J.P. Nossin of ERLI), was aimed at providing a common model able to express the different types of lexical information owned by the partners (publishers, research institutes, language industry providers, etc.) and a set of software tools to insert this information into the model and translate it into a well-defined formalism.

MULTILEX (ESPRIT – industry, co-ordinated by Katchaturion of CAP GEMINI) aimed at defining a common format for encoding multilingual lexicons.

All the co-ordinators of these projects, presenting their work programme at the first MULTILEX meeting (Paris, 1991), listed among their goals the definition of technical specifications for building computational lexical entries, each stressing that their specifications had been proposed to function as 'standards' for the R&D community.

'Shocked' by the contradiction implicit in the fact that four different projects, all supported by the EU, were aiming at proposing 'European standards' for the same class of lexical objects, and 'afraid' of the risk of duplication and intellectual and financial efforts, I invited the 4 co-ordinators to a meeting in Pisa. There we decided to ask for the support of the Commission in establishing co-ordination among the various projects, and thus to joint efforts concerning the proposal of a common standard.

The Commission accepted our proposal, sponsoring a 'Co-ordination Group of the Lexical Projects', which was later extended to form EAGLES (Expert Advisory Group on Language Engineering Standards).

The EAGLES initiative aims at accelerating the provision of standards for:

(i) very large-scale language resources (such as text corpora, computational lexicons and speech corpora);

(ii) means of manipulating such knowledge, via computational linguistic formalisms, mark-up languages and various software tools;

(iii) means of assessing and evaluating resources, tools and products.

Leading industrial and academic elements in the Language Engineering field have actively participated in the definition of this initiative and have lent invaluable support to its achievement. Moreover, the initiative is a direct result of a series of recommendations made to the EU over several years. Reports from EU language engineering strategy committees have strongly endorsed standardisation efforts in language engineering. The mid-term review of the EU's Telematics Programme of July 1993 states:

> The importance of working towards standards and protocols is well recognised with the establishment of the EAGLES project that brings together senior representatives of all the major speech and language development projects in Europe.
>
> (Oakley 1993, personal communication).

Moreover, there is a recognition that standardisation work is not only important, but is a necessary component of any strategic programme to promote the advancement of the current technology and to create a coherent market, which demands sustained effort and investment.

The EAGLES initiative is run by an organisational structure similar – to a certain extent – to theat of the TEI:

- a Co-ordination Team in Pisa, including the co-ordinator of the project, the project editors;
- a Management Board, formed by representatives of major Associations, projects, companies in the field;
- a Working Group (WG) for each of the major topics, co-ordinated by a WG chair, assisted by a WG editor (or editorial team);

- a Technical Committee, formed by the co-ordinator, the editors of the project, the chairman and the editors of the WGs.

In the first two phases of EAGLES, the following WGs were activated:

- 1st phase (1995-97): corpora, lexica (syntax), formal grammars, speech technologies, evaluation;
- 2nd phase lexica (1997-99) (semantics), speech (including linguistically annotated dialogues), evaluation.

In the last phase (EAGLES – ISLE: **I**nternational **S**tandards for **L**anguage **E**ngineering) (2000-2002), the workplan envisages three major items and therefore three WGs: (i) multilingual lexicons, (ii) natural interaction and multimodality (NIMM), and (iii) evaluation of HLT systems. These areas were chosen not only for their relevance to the current HLT call but also for their long-term significance. For multilingual computational lexicons, ISLE will: extend EAGLES work on lexical semantics, necessary to establish inter-language links; design standards for multilingual lexicons; develop a prototype tool to implement lexicon guidelines and standards; create exemplary EAGLES-conformant sample lexicons and tagged exemplary corpora for validation purposes; develop standardised evaluation procedures for lexicons. For NIMM, a rapidly innovating domain urgently requiring early standardisation, ISLE will develop guidelines for the creation of NIMM data resources; interpretative annotation of NIMM data, including spoken dialogue in NIMM contexts; annotation of discourse phenomena. For evaluation, ISLE will work on quality models for machine translation systems and maintenance of previous guidelines – in an ISO based framework (ISO 9126, ISO 14598).

We might say that EAGLES is now a well-established brand name, and that the recommendations are widely used, particularly in certain areas. For example, the EAGLES proposals for the morphosyntactic level are adopted in more than 300 sites for more than 20 languages.

the linguistic phenomena which should be encoded, more than on their formal representation.

### *2.4 Concluding remarks*

Promotion, dissemination, discussion, feedback and continuous interaction with users are a central concern of both the TEI and EAGLES.

All these, plus, in particular, maintenance and updating of the Guidelines and Recommendations, require the presence of a permanent structure, which could ensure the necessary continuity.

Unfortunately, the Funding Agencies (as the EC, the NSF, the NEH) do not normally fund infrastructural costs; they can only support research projects of limited duration.

This, of course, has created a serious problem for EAGLES, TEI and in general all the activities whose results would be lost without appropriate management of feedback and continuous updating and maintenance. Infrastructural LR is a clear example.

The TEI is now in the process of seeking a solution to this problem, through a Consortium of TEI users and developers which could ensure the necessary continuity and basic infrastructure.

I believe that standardisation efforts should be extended to other scientific disciplines, which should develop guidelines for encoding their analytical-interpretative categories.

The new field of multimodal and multimedia resources seems to offer a natural area of convergence and co-operation of various humanistic disciplines, social sciences and HLT.

It is important to note that – in a number of discussions between European and North American experts – standards for LR have been recognised as a priority for the implementation of the recently signed Transatlantic Scientific and Technical Co-operation Agreement ISLE/EAGLES is jointly supported by the EU and the National Science Foundation (NSF) programmes.

### 2.2.2 TEI

In 1988, Nancy Ide (Vassar College, USA), sensing the need to conclude the already decade-long discussion on standards for humanities texts representation, organised a Workshop at Vassar . Following the example of the Grosseto Workshop, ALLC, ACH, and ACL sponsored the meeting.

About 30 representatives of specialised centres, textual archives, digital projects, were convened to discuss the desirability and feasibility of a common encoding scheme/format for intercharging texts in MRF for HTP. The discussion clearly indicated that such interchange format was a common desire.

In order to avoid the risk that the general consensus reached at the end of the discussion would remain a dead letter without a practical follow-up, clear action was needed immediately: I called the chairs of the three Associations at a post-dinner meeting the last night in Vassar, inviting them to a preparatory organisational meeting in Pisa. There, the three Associations agreed on jointly launching the 'Text Encoding Initiative'. Practical actions to be taken in order to establish the organisational structure, prepare the workplan, and find the necessary funding, were also agreed.

A Steering Committee of six people was created by the three Associations which appointed 2 members each (ALLC: Susan Hockey and Antonio Zampolli; ACL: Don Walker and Robert Amsler; ACH: Nancy Ide and Michael Sperberg McQueen).

In setting up the organisational and operational structure, the Steering Committee tried, as far as possible, to ensure an equal participation of American and European experts in each component.

The Steering Committee appointed the two editors (one European: Lou Burnard; one American: Michael Sperberg McQueen); designed an initial overall workplan, appointing four WGs (members and chairs) to perform it: text documentation (M. Sperberg McQueen), text representation (S. Johansson), text analysis and interpretation (T. Langendoen), metalanguage and syntax (D.T. Barnard); organised the Advisory Board, inviting various scholarly associations to appoint a representative; it constantly ensured the scientific, organisational, financial, operational monitoring of the work.

ACH, through the University of Chicago at Illinois, applied to the NEH in order to obtain the funding for American participation (about $800.000 ).

I was able to obtain the support for European participation, acting on behalf of ALLC, in the form of four successive contracts between EC and the University of Pisa (for a total of about 600.000 ECUs ).

ACL was able to obtain a grant from the Mellon Foundation.

It would be interesting to identify which factors made it possible, finally, to launch the TEI. Having personally gone through the process, I have identified the following items:

- the convergence of previous efforts;
- the awareness promoted in major Funding Agencies – national and international – of the need for standards (in particular the ESF and the Grosseto Workshops);
- the advent of the Language Industry Paradigm, which favoured synergy between NLP and HTP, and drew attention to the central role of LR;
- the co-operation of the three major Associations (ALLC, ACH, ACL);
- the diffusion of PCs: a large number of texts in MRF were/are created by isolated researchers (who need guidelines for encoding, whereas large text centres have their own traditional encoding practices;

- the technical context offered by telematic networking: interest in portability, protocols, etc.;
- SGML availability;
- the success of digital libraries;
- the maturity of text processing methods and the variety of text types to which they are applied;
- the increasing number of texts available in MRF;
- the personal engagement and commitment of the Steering Committee members at a technical, scientific, organisational level.

Initial start-up funds up were made available by the Institute of Computational Linguistics of the National Research Council and by the University of Pisa, motivated by the consideration that the provision of standards for the HTP community was part of their institutional mandate.

Thanks to their contacts and knowledge in the field, the Steering Committee members were able to involve large, well-established text centres and archives, identifying potentially interested funding sources, and obtaining substantial financial support with the appropriate motivations.

### *2.3. A comparison of EAGLES and TEI*

So far we have pointed out analogies and similarities between the TEI and EAGLES, underlying that both were originated by researchers capable of clearly identifying needs and interpreting demands emerging from the R&D community, shaping organisational structures apt to carry out the necessary work, and obtaining the financial support of various Funding Agencies.

We will now consider some major differences in the scope of the two initiatives.

### 2.3.1. Some major differences between EAGLES and TEI

#### *2.3.1.1 Scope and goal of TEI*

According to the P3 Guidelines (Sperberg McQueen / Burnard 1994: 3), the central goal of the TEI is to serve the community of humanistic researchers, providing (i) a common format for text interchange, and (ii) a guide to encode texts in this format, in particular to encode all the features which should be explicitly marked in order to allow/facilitate HTP. Without these markers, many important features would be difficult/impossible to be automatically recognised.

The Guidelines provide an inventory of features which, according to the experiences of the community, are useful for HTP, thus promoting the reuse of a text in MRF for a plurality of research activities, and providing guidance to newcomers in the field of text encoding.

The inventory of these features depends on the needs of the process that uses them. HTP software today usually includes components which offer the following functionality:

- to select individual texts in an electronic library (bibliographical and situational data);
- to select specific part of texts (structural markers, highlighted syntagms, etc.);
- to 'compute' frequencies, patterns, contexts, etc., of graphical sequences.

For this reason, the 'core tagset' of the TEI essentially includes information traditionally represented by bibliographical and typographical practices.

#### *2.3.1.2 Scope of EAGLES*

The overall scope of EAGLES is to serve research and development in the field of language engineering/human language technology/language industry, in particular providing the opportunity to build reusable LR in order to ensure:

- availability of the necessary linguistic infrastructure for as many languages as possible;
- economy of scale (costs, times, efforts), promoting co-operation and avoiding duplications in the creation of the LR;
- the concentration of efforts of R&D on new areas made possible by the consolidation of acquired knowledge and results.

To this end, EAGLES tries to identify areas for which a consensus does not exist yet but which are mature enough to promote it through the interactions of the major experts/elements.

As an example of the different approaches take by EAGLES and TEI, we might consider the case of 'Dictionaries'.

The TEI P3 (p. 321) says: 'The chapter on dictionaries defines a set of basic tagsets to encode human-oriented Dictionaries, as opposed to computational lexica, which are used by NLP software'.

Typically, a computational lexicon should , for each reading of a lexical entry, provide the information required by the parser/semantic interpreter/inference mechanism, etc., such as: part of speech, argument structure, syntactical form of the arguments, formal semantic characterisation of the meaning, relations to other meanings (for ex: synonymy, hyponymy, metonymy, connection to an ontology), domain, pragmatic/conceptual field, etc.

The different NLP components, and the different linguistic theories behind them, use different categories, properties, formalisms, etc.. EAGLES aims at providing a 'polytheoretical', 'multifunctional' identification and representation, from which each theory or system could automatically derive the information in the format it needs.

### 2.3.2 Interpretation versus representation

The TEI Guidelines make a distinction between 'objective' (representational) and 'subjective' (interpretative) features/infor–

mation. According to the Guidelines (p. 6), this distinction can be seen as a distinction between topics on which consensus exists/does not exist. The Guidelines say that the distribution of items to be dealt with, respectively, in the "Text Representation" and in the "Text Interpretation and Analysis" Committees, was due 'to contingent factors and not to an agreed definition'.

It is true that the borderline between 'representational' and 'interpretative' features is in some cases blurred. But I think that, essentially, the 'interpretative' features are the ones which encode the results of the analysis which a specific discipline typically performs on a text, and the knowledge of this discipline about the relevant properties of its units of analysis.

Consider simple linguistic examples: units of analysis can be syntactic phrases, accompanied by the description of their classification (nominal, verbal, adjectival groups, etc.), functions (object, subject, predicate, etc.), relations (their recursive combinations in larger syntactic units), etc.

Typographical features are extremely important also for NLP. For example, a MT programme should maintain, producing a translation, the typical features of the source text. The 'grammar' of titles can describe different regularities than the grammar of texts, etc..

But, typically, NLP does not operate, or not only, on graphical forms, but rather on linguistic, conceptual or 'meaning' units and their description. In general, before performing its own specific applicative task (translation, summarisation, extraction, etc.), a NLP system aims at automatically recognising linguistic units of certain level(s (morphological, syntactical, semantic, etc.).

This recognition requires the consultation of repositories of linguistic information and descriptions – usually in the form of computational lexicons.

According to a well-known practice, if not a new paradigm, it is increasingl more common to see the various components of an NLP system 'learn' from, or 'are trained' on large and linguistically annotated, textual corpora.

### 2.3.3 Method

As mentioned above, EAGLES aims to contribute to the feasibility of reusable (multifunctional and polytheoretical) LR, trying to establish, for areas of sufficient maturity, a consensus among the researchers/developers in the field. For example, on the typology and organisation of the linguistic information to be encoded in the lexicon, so that various parsers can (easily) derive (possibly through an automatic conversion process) the relevant information in the specific form they require.

To this end EAGLES adopts an explicit working methodology, which has the following major steps:

- survey and inventory phase;
- discussion phase in WG meetings to achieve consensus;
- drafting of preliminary recommendations;
- validation actions for testing the practical applicability of the proposals (e.g. through preparation of small test resources);
- external evaluation: User Group, other projects, external experts, other languages;
- integration of feedback;
- definition of formal specifications and operational guidelines;
- final recommendations.

I believe that, at least in principle, TEI has adopted a methodology including – albeit less explicitly – very similar steps. The difference is:

- TEI tries to establish consensus on the formal representation of textual features on whose inventory classification and definition a consensus, normally channelled by the typographical tradition, already exists;
- EAGLES tries to build or make explicit the consensus of the community on the identification, classification, definition of