Chapter 1

LMF – Historical Context and Perspectives

1.1. Introduction

The value of agreeing on standards for lexical resources was first recognized in the 1980s, with the pioneering initiatives in the field of machine-readable dictionaries, and afterwards with EC-sponsored projects ACQUILEX, MULTILEX and GENELEX. Later on, the importance of designing standards for language resources (LR) was firmly established, starting with the Expert Advisory Group for Language Engineering (EAGLES) and International Standards for Language Engineering (ISLE) initiatives. EAGLES drew inspiration from the results of previous major projects, set up the basic methodological principles for standardization and contributed to advancing the common understanding of harmonization issues. ISLE consolidated the uncontroversial basic notion of a lexical metamodel, that is an abstract representation format for lexical entries, the Multilingual ISLE Lexical Entry (MILE). MILE was a general schema for the encoding of multilingual lexical information, and was intended as a common representational layer for multilingual lexical resources. As such, all these initiatives contain the seeds of what later evolved into Lexical Markup Framework (LMF). From a methodological point of view, MILE was based on a very extended survey of common practices in lexical encoding, and was the result of cooperative work toward a consensual view, carried out by several groups of experts worldwide. Both EAGLES and

---

Chapter written by Nicoletta CALZOLARI, Monica MONACHINI and Claudia SORIA.

ISLE stressed the importance of reaching a consensus on (linguistic and non-linguistic) "content", in addition to agreement on formats and encoding issues, and also began to address the needs of content processing and Semantic Web technologies. The recommendations for standards and best practices issued within these projects then became, through the INTERA and mainly the LIRICS project, the International Organization for Standardization (ISO) within the ISO TC37/SC4 committee, where LMF was developed. Thanks to the results of these initiatives that culminated in LMF, there is worldwide recognition that the EU is at the forefront in the areas of LRs and standards. LMF now testifies the full maturity reached by the field of LRs.

1.2. The context

The 1990s saw a widespread acknowledgment of the crucial role covered by LRs in language technology (LT). LR started to be considered as having an infrastructural role, that is as an enabling component of Human Language Technologies (HLTs). HLTs (i.e. natural language processing tools, systems, applications and evaluations) depend on LRs, which also strongly influence their quality and indirectly generate value for producers and users.

This recognition was also shown through the financial support from the European Commission to projects aiming at designing and building different types of LRs. Under the support of US agencies (NSF, DARPA, NSA, etc.) and the EC, LRs were unanimously indicated as themes of utmost priority.

One of the major tenets was the recognition of the essential infrastructural role that LRs play as the necessary common platform on which new technologies and applications must be based. To avoid massive and wasteful duplication of effort, public funding – at least partially – of LR development is critical to ensure public availability (although not necessarily at no cost). A prerequisite to such a publicly funded effort is careful consideration of the needs of the community, in particular the needs of industry. In a multilingual setting such as today's global economy, the need for standardized wide-coverage LRs is even stronger. Another tenet is the recognition of the need for a global strategic vision, encompassing different types of (and methodologies of building) LR, for an articulated and coherent development of this field.

The infrastructural role of LRs requires that they are (1) designed, built and validated together with potential users (therefore, the need for involving companies), (2) built reusing available "partial" resources, (3) made available to the whole community and (4) harmonized with the resources of other languages (therefore, the importance and the reference to international standards).

The major building blocks to set up an LR infrastructure are presented in [CAL 99]:

– *LR reusability*: directly related to the importance of "large-scale" LRs within the increasingly dominant data-driven approach;

– LR development;

– LR distribution.

Other dimensions were soon added as a necessary complement to achieve the required robustness and data coverage and to assess results obtained with current methodologies and techniques, that is:

– automatic acquisition of LRs or of linguistic information;

– use of LRs for evaluation campaigns.

Crucial to LR reusability and development was the theme of the definition of operational standards, but the value of agreeing on International Standards was also suddenly recognized as critical. Without standards underlying applications and resources, users of LT would have remained ill-served. The application areas would have continued to be severely hampered and only niche or highly specialized applications would have seen success (e.g. speech aids for the disabled and spelling checkers). In general, it had never been possible to build on the results of past work, whether in terms of resources or the systems that used them.

The significance of standardization was thus recognized, in that it would open up the application field, allow an expansion of activities, sharing of expensive resources, reuse of components and rapid construction of integrated, robust, multilingual language processing environments for end-users.

1.3. The foundations: the Grosseto Workshop and the "*X-Lex*" projects

During the 1980s there was a dramatic growth in interest in the lexicon. The main reasons for this were, on the one hand, the theoretical developments in linguistics that placed increasing emphasis on the lexical component, and on the other hand the awareness about the wealth of information in lexicons that could be exploited by automatic NLP systems. A turning point in the field was marked by the workshop "On automating the lexicon" held at Marina di Grosseto (Italy) in 1986 [WAL 95], when a pool of actors in the field gathered to establish a baseline for the current state of research and issued a set of recommendations for the sector. The most relevant recommendation – as far as the future LMF is concerned – was the need for a metaformat for the representation of lexical entries, that is an abstract model of a computerized lexicon enabling accommodation of different theories and linguistic models. The following years saw a flourishing of events around this new notion of a "meta-entry", for instance the workshop on "The Lexical Entry", held in New York City immediately after Grosseto, and the meeting held in Pisa by the so-called Polytheoretical Group in 1987, where the possibilities of a neutral lexicon were explored [WAL 87].

This has contributed to the creation of a favorable climate for converging toward the common goal of demonstrating the feasibility of large lexicons, which needed to be *reusable*, *polytheoretical* and *multifunctional*. This reflection has led to the definition of the concept of reusability of lexical resources as (1) the possibility of reusing the wealth of information contained in machine-readable dictionaries, by converting their data for incorporation into a variety of different NLP modules; (2) the feasibility of building large-scale lexical resources that can be reused in different theoretical frameworks, for different types of application, and by different users [CAL 91].

The first sense of reusability was clearly addressed by the ACQUILEX project, funded by the European ESPRIT Basic Research Program [BOG 88]. The second sense inspired the Eurotra-7 (ET-7) project, which had the goal of providing a methodology and recommending steps toward the construction of sharable lexical resources [HEI 91].

The need for standards in the second sense of reusability was represented by other initiatives, often publicly funded, such as the EUREKA industrial

project GENELEX [GEN 94], which concentrated on a generic model for monolingual reusable lexicons [ANT 94] and the CEC ESPRIT project MULTILEX, whose objective was to devise a model for multilingual lexicons [KHA 93]. GENELEX, with its generic model, fulfilled the requirements of being "theory welcoming", and having a wide linguistic coverage. A standardized format was designed as a means for encoding information originating from different lexicographic theories, with the aim to make it possible to exchange lexical data and to allow the development of a set of tools for a lexicographic workstation.

These "X-Lex" projects assessed the feasibility of some elementary standards for the description of lexical entries at different levels of linguistic description (phonetic, phonological, etc.) and laid the foundations for all the subsequent standardization initiatives.

It became evident that progress in NLP and speech applications were hampered by a lack of generic technologies and reusable LRs, by a proliferation of different information formats, by variable linguistic specificity of existing information and by the high cost of development of resources. This had to be changed to be able to build on the results of past work, whether in terms of resources or the systems that use them.

1.4. EAGLES and ISLE

EAGLES, started in 1993, is a direct descendant of the previous initiatives, and represented the bridge between them and a number of subsequent projects funded by the EC [CAL 96]. EAGLES was set up to improve the situation of many lexical initiatives, through bringing together representatives of major collaborative European R&D projects in relevant areas, to determine which aspects of our field are open to short-term *de facto* standardization and to encourage the development of such standards for the benefit of consumers and producers of LT. This work was conducted with a view to providing the foundation for any future recommendations for International Standards that may be formulated under the aegis of ISO.

The aim of EAGLES was to support academic and industrial research and development in HLT by accelerating the provision of standards, common guidelines and best practice recommendations for:

– very large-scale LRs (such as text corpora, computational lexicons and speech and multimodal resources);

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

– means of assessing and evaluating resources, tools and products.

The structure of EAGLES resulted from recommendations made by leading industrial and academic centers, and by the EC Language Engineering strategy committees. More than 30 research centers, industrial organizations, professional associations and networks across the EU provided labor toward the common effort, and more than 100 sites were involved in different EAGLES groups or subgroups. In addition, reports from EC Language Engineering strategy committees had strongly endorsed standardization efforts in language engineering.

Moreover, there was a recognition that standardization work is not only important, but is also a necessary component of any strategic program to create a coherent market, which demands sustained effort and investment. ISLE, a standard-oriented transatlantic initiative under the HLT program, started in 2000, was a continuation of the long-standing European EAGLES initiative [CAL 01, CAL 02].

It is important to note that the work of EAGLES/ISLE must be seen in a long-term perspective. This is especially true for any attempt aiming at standardization in terms of international standards. EAGLES did not and could not result in standards of such an impact: this is the preserve of the ISO. The basic idea behind EAGLES/ISLE work was for the group to act as a catalyst in order to pool concrete results coming from major international/national/industrial projects.

1.5. Setting up methodologies and principles for standards

From a retrospective point of view, it is important to note that EAGLES and its guidelines were the first attempt at defining standards directly responding to commonly perceived needs in order to overcome common

problems. In terms of offering workable, compromise solutions, they must be based on a solid platform of accepted facts and acceptable practices.

Since the formation of EAGLES, the work related to standards in the EU has largely been concentrated within this initiative. Related efforts elsewhere were closely linked with EAGLES and feed off it. The Lexicon and Corpus groups' recommendations were soon applied in a large number of European and national projects. Indeed, EAGLES has acted as a catalyst and testing ground.

EAGLES drew strong inspiration from the results of major projects whose results had contributed to advancing our understanding of harmonization issues. Relevant common practices or upcoming standards were used where appropriate as input to EAGLES/ISLE work. Several LRE projects have been active in contributing comments and in testing EAGLES proposals, thus offering a concrete industry-related setting. Given the amount of industrial participation in EAGLES itself, it is notable that there has been significant advances in Language Engineering Standards, thus re-emphasizing the need to involve industry in such efforts in targeting clearly identified and motivated standardization goals. EAGLES results are to be seen as a first step on the path toward standardization for language engineering purposes.

The major efforts in EAGLES concentrate on the following types of activities:

– detecting those areas ripe for short-term standardization versus areas still in need of basic research and development [EAG 96b];

– assessing and discovering areas where there is a consensus across existing linguistic resources, formalisms and common practices;

– surveying and assessing available proposals or contributed specifications in order to evaluate the potential for harmonization and convergence and for the emergence of standards;

– proposing common specifications for core sets of basic phenomena, recommendations for good practice, for standard methodologies, etc. on which a consensus can be found [MON 96];

– setting up guidelines for the representation of core sets of basic features, for the representation of resources, etc. [LEE 96];

– collecting and cataloging information on spoken LRs and *de facto* standard procedures, and providing an essential reference work for speech technology development [GIB 97];

– carrying out feasibility studies for less mature areas [EAG 99];

– suggesting actions to be taken for a stepwise procedure leading to the creation of multilingual reusable resources, elaboration of evaluation methodologies and tools [EAG 96a], etc.

This method of work has proven useful in the process of reaching consensual *de facto* standards in a bottom-up approach and was also at the basis of ISLE work.

1.5.1. *The MILE methodology: toward LMF*

The new awareness created by EAGLES regarding the need to reconcile different approaches to LR building was the direct inspiration for the new concept of "edited union". This term, coined by Gerald Gazdar in one of the first EAGLES meetings, refers to the idea of conciliating what exists in major lexicons/models/dictionaries. This concept shaped the *MILE*, that is a highly *modular* and *layered* structure, with different levels of recommendations [BER 04]. The MILE was intended as a meta-entry, acting as a common representational layer for multilingual lexical resources. The key ideas underlying the design of a meta-entry can be summarized as follows. Different theoretical frameworks appear to impose different requirements on how lexical information should be represented. One way of tackling the issue of theoretical compatibility stems from the observation that existing representational frameworks mostly differ in the way pieces of linguistic information are mutually implied, rather than in the intrinsic nature of this information.

MILE is the direct ancestor of LMF. We will not describe MILE in detail here, but we will just introduce some of the basic notions at the basis of MILE, because these notions are also important for LMF.

The MILE was designed to meet the following desiderata:

– factor out linguistically independent (but possibly correlated) primitive units of lexical information;

– make information explicit and accessible by NLP systems;

– rely on lexical analysis that have the highest degree of inter-theoretical agreement;

– avoid framework-specific representational solutions.

All these requirements served the main purpose of making the lexical meta-entry open to task- and system-dependent parameterization.

The MILE lexicon architecture built, in particular, on the results of the EUREKA GENELEX and the ESPRIT MULTILEX projects, to design a multilingual and multifunctional lexicon model. Such architecture embodied three levels of linguistic information: obligatory, recommended and optional (optional splits furthermore into language independent and language dependent). In this way, the MILE modularity addressed three basic principles: (1) flexibility of the representation, (2) easiness of customization and integration of existing resources and (3) usability by different systems which may need different portions of the data.

The descriptive granularity of the MILE aimed at reaching a maximal *decomposition* into minimal basic information units. Therefore, small units can be assembled, in different frameworks, according to different (theory/application dependent) generalization principles. For instance, the MILE allowed us to decompose a theory-specific complex notion, such as "synset", into theory-neutral minimal basic units, such as "senses", "semantic relations", where "synonymy" is a particular instance of semantic relation.

On the other side, past EAGLES experience had shown that it was useful in many cases to accept *underspecification* with respect to recommendations for the representation of some phenomenon (and *hierarchical structure* of the basic notions, attributes, values, etc.): (1) to allow for agreement on a minimal level of specificity especially in cases where we cannot reach wider agreement and/or (2) enable mapping and comparability of different lexicons, with different granularity, at the minimal common level of specificity (or maximal generality). For example, the work on syntactic subcategorization in EAGLES proved that it was problematic to reach agreement on a few notions, for example it seemed unrealistic to agree on a set of grammatical functions. This has led to an underspecified recommendation, but nevertheless recommendation that was useful.

Another key strategy adopted was the continuous, *cyclic interaction* between EAGLES and a large number of topic-specific R&D projects and applications.

1.6. EAGLES/ISLE legacy

EAGLES/ISLE, thus, was very influential for the field in providing the mold that shaped the representation of LRs for the years to come. Its heritage gave rise to a burning activity in the development and annotation of LRs, and directly informed the work later on carried out within the ISO Committee devoted to Language Resource Management and Representation. Beside this theoretical legacy, the other main achievement of EAGLES/ISLE was that it provided cohesion to the community engaged in the LR and technology sector.

We identify at least three main footprints. The first two refer to *low-level specifications*, that is recommendations related to the linguistic categories used for linguistic representation. The third refers to an abstract representation level, as a set of high-level objects used for describing the structural components of LRs.

First, *a common core of morphosyntactic distinctions* to be encoded in corpora and lexicons. Comparison of how morphosyntactic phenomena are encoded for all EU languages has led to a proposal for encoding a common core of morphosyntactic distinctions in a multilayered structure with applications for all the EU languages (also Eastern Europe), which gives the user more flexibility thus (1) allowing him/her to choose the most appropriate level of granularity and (2) providing a straightforward framework for extensions and updating. These specifications represent the basis on which the data categories of the ISO-12620 were developed within the morphosyntactic Thematic Domain Group, and now embodied in ISOCat.

Second, *a common approach to subcategorization in syntax*. Comparison of how different systems and theories in different European languages classify and deal with subcategorization phenomena has led to a preliminary classificatory scheme and to the proposal of a set of standardized basic notions for subcategorization, using a frame-based structure.

The EAGLES morphosyntactic guidelines [MON 96, LEE 96] were applied – and consequently tested and evaluated – in the LE-PAROLE Project for the syntactic layer of 12 EU languages, and in a very large number of other national and European projects, such as LRE DELIS, RENOS, CRATER, MECOLB, MULTEXT, COPERNICUS MULTEXT-East and TELRI, MLAP-PAROLE, ESPRIT-ELSNET, French GRACE, German Textcorpora *und Erschliessungswerkzeuge*, LE-SPARKLE, EUROWORDNET and Italian national projects.

Third, the *provision of a proposal for a multilingual and multifunctional model for a lexicon*, viewed as a resource out of which to extract specific application lexicons.

EAGLES results in many areas, through their application in numerous projects, became *de facto* widely adopted standards, and became a well-known trademark and a point of reference for HLT projects and products. EAGLES work toward *de facto* standards allowed the field of LRs to establish a broad consensus on key issues for some well-established areas, thus providing a key opportunity for further consolidation and a basis for technological advance.

The idea of a standard model for lexicon architecture originated here: the LMF [FRA 06] standard adopts a modular organization to cope with the challenge that actual lexicons differ very much both in complexity and type of encoded information. LMF is made up of a core model, a sort of simple skeleton and various semi-independent packages of notions, used for the various linguistic layers that make up a lexicon.

We wish to highlight here the importance of having both a standard model and core LRs (e.g. corpora and lexicons) also encoded according to the standard – or even more – for applications in the humanities. It may be in fact a big advantage to have the possibility of referring to and adopting available guidelines and possibly reusing available harmonized LRs, thus concentrating research efforts on issues more pertinent to the specific field of interest.

EAGLES results in the Lexicon and Corpus areas were adopted by an impressive number of European – and also national – projects, thus becoming "the *de-facto* standard" for LR in Europe. This is a very good measure of the impact – and of the need – of such a standardization initiative

in the HLT sector. To mention just a few key examples, the LE PAROLE/SIMPLE resources (morphological/syntactic/semantic lexicons and corpora for 12 EU languages) [RUI 98, LEN 99, BEL 00] rely on EAGLES results [EAG 96b, EAG 99], and were then enlarged at the national level through many national projects. The fact that the core PAROLE/SIMPLE resources were enlarged to real-size lexicons within national projects in at least eight EU countries was a big step toward a very large infrastructural platform of harmonized lexicons in Europe, sharing the same model. Moreover, the ELRA Validation Manuals for Lexicons [UND 97] and Corpora [BUR 97] are based on EAGLES guidelines.

1.6.1. *Lessons learned for standard design*

From a retrospective point of view, the experience gained in those years was influential, in particular from the point of view of the leading principles that must guide the standardization process. Standards must emerge from state-of-the-art developments and as such they are not to be imposed. Consolidation of a standard's proposal must be viewed, by necessity, as a slow process and, by definition, as a non-innovative action. The process of standardization, although by its own nature not intrinsically innovative, must – and actually does – proceed shoulder to shoulder with the most advanced research. Since EAGLES involved many bodies active in EU–US NLP and speech projects, close collaboration with these projects was assured and, significantly, in many cases, free manpower has been contributed by the projects, which is a sign of both the commitment of these groups/companies and of the crucial importance they place on reusability issues.

After the phase of putting proposals forward, it must comprise a cyclical phase involving external groups and projects with:

– careful evaluation and testing by the scientific community of recommendations in concrete applications;

– application, if appropriate, to a large number of languages;

– feedback on and readjustment of the proposals until a stable platform is reached, upon which a real consensus – acquiring its meaning by real usage – is achieved;

– dissemination and promotion of consensual proposals.

This long process has the merit of making new areas for consensus emerge while promoting consciousness of their stability in the community at the same time.

Finally, one of the targets of standardization is to create a common parlance among the various actors (both of the scientific and the industrial R&D community) in the field of computational lexical semantics and multilingual lexicons, so that synergies will be enhanced, commonalities strengthened and resources and findings usefully shared. In other terms, the process of standard definition undertaken by EAGLES, and by the ISLE enterprise in particular, represents an essential interface between advanced research in the field of multilingual lexical semantics and the practical task of developing resources for HLT systems and applications. It is through this interface that the crucial trade-off between research practice and applicative needs can actually be achieved.

1.6.2. *Moving closer to LMF*

After the EAGLES/ISLE experience, and the subsequent use of their results in so many projects, the ground was ready to move from standards and best practices directly emerging from projects and research groups to an international, coordinated and structured effort ratified by standardization organizations. A new work item proposal was issued by the ISO/TC37 US delegation in Summer 2003. In Fall 2003, the French delegation issued a technical proposition for a data model dedicated to NLP lexicons. In early 2004, the ISO/TC37 committee decided to form a common ISO project with Nicoletta Calzolari (CNR-ILC Italy) as convenor and Gil Francopoulo (Tagmatica France) and Monte George (ANSI USA) as editors. This was the start of the LMF (ISO-24613). From 2005 to 2007, the ISO activities were carried out in parallel with the EU eContent project LIRICS (http://lirics.loria.fr).

The goals of this project were to provide ISO ratified standards for LT to enable the exchange and reuse of multilingual LRs, and at the same time to facilitate the implementation of these standards for end-users. Through an Industry Advisory Group and demonstration workshops, LIRICS managed to gain full industry support and input for the standard's development. The LIRICS Consortium brought together leading experts in the field of NLP and related standards development via participation in ISO committees and

National Standardization committees, closely following the procedures established by ISO.

The first step in developing LMF was to design an overall framework based on the general features of existing lexicons and to develop a consistent terminology to describe the components of those lexicons. The following step was the actual design of a comprehensive model that best represented all of the lexicons in detail. A large panel of 60 experts contributed a wide range of requirements for LMF that covered many types of NLP lexicon. The editors of LMF worked closely with the panel of experts to identify the best solutions and reach a consensus on the design of LMF. Special attention was paid to the morphology in order to provide powerful mechanisms for handling problems in several languages that were known as difficult to handle. A total of 13 versions have been written, dispatched (to the national nominated experts), commented upon and discussed during various ISO technical meetings. After 5 years of work, the editors arrived at a coherent UML model. In conclusion, LMF should be considered a synthesis of the state of the art in NLP lexicon field.

1.7. Interoperability: the keystone of the field

Since the first attempts, and after LMF, we have made big steps forward with respect to *interoperability*. Today, open, collaborative, shared data are at the core of a sound language strategy. Standards are fundamental to exchange, preserve, maintain and integrate data and LRs, to achieve interoperability in general, and they are an essential basis of any LR infrastructure.

What was called “reusability” in the past has evolved today into “interoperability”. Interoperability means the ability of information and communication systems to exchange data and to enable the sharing of information and knowledge. To make the notion of interoperability operational, we need to set up an interoperability framework. This can be described as a dynamic environment of language (and other) standards and guidelines, where different standards are coherently related to one another and guidelines clearly describe how the specifications may be applied to various types of resource. Such a framework should be internally coherent, that is a series of specific standards should continue to exist, but they should form a coherent system (i.e. coherence among the various standard

specifications must be ensured so that they can "speak" to each other). The framework should also be dynamic, in the sense that standards must be conceived as dynamic, because they need to follow and adapt to new technologies and domains of application. As the LT field is expanding, standards need to be periodically revised, updated and integrated in order to keep pace with technological advancement.

An interoperability framework is also intended to support the provision of language service interoperability. Enterprises nowadays seem to need such a language strategy, and to be key players they must rely on interoperability, otherwise they are out of business. A recent report by TAUS [TAU 11] states that: "The lack of interoperability costs the translation industry a fortune", where the highest price is paid mainly for adjusting data formats.

The community and funding agencies need to join forces to drive forward the use of existing and emerging standards, at least in the areas where there is some degree of consensus. The only way to ensure useful feedback to improve and advance is to use these standards on a regular basis. It will thus be even more important to enforce and promote the use of standards at all stages, from basic standardization for less-resourced languages (such as orthography normalization and transcription of oral data) to more complex areas (such as syntax and semantics).

However, enforcing standards cannot be a purely top-down process. It must be backed by information about contributions from different user communities. As most users are not very concerned about whether or not they are using standards, there should be easy-to-use tools that help them apply standards while hiding most of the technicalities. The goal would be to have standards operating in the background as "intrinsic" properties of the LT or the more generic tools that people/end-users use.

But true content interoperability is still far away. We may have solved the issue of formats, of inventories of linguistic categories for the various linguistic layers, but we have not solved the problem of relating senses, which would allow automatic integration of semantic resources. This is a challenge for the following years, and a prerequisite for both a true Lexical Web and a credible Semantic Web.

1.8. Bibliography

[ANT 94] ANTONI-LAY M.H., FRANCOPOULO G., ZAYSSER L., "A generic model for reusable lexicons: the genelex project", in OSTLER N., ZAMPOLLI A. (eds), *Literary and Linguistic Computing*, vol. 9, no. 1, pp. 47–54, 1994.

[BEL 00] BEL N., BUSA F., CALZOLARI N., GOLA E., LENCI A., MONACHINI M., OGONOWSKI A., PETERS I., PETERS W., RUIMY N., VILLEGAS M., ZAMPOLLI A., "SIMPLE: a general framework for the development of multilingual lexicons", *LREC Proceedings*, Athens, 2000.

[BER 04] BERTAGNA F., LENCI A., MONACHINI M., CALZOLARI N., "Content interoperability of lexical resources: open issues and 'MILE' perspectives", *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, ELRA, 2004.

[BOG 88] BOGURAEV B., BRISCOE E.J., CALZOLARI N., CATER A., MEIJS W., ZAMPOLLI A., Acquisition of lexical knowledge for natural language processing systems (ACQUILEX), Proposal for ESPRIT Basic Research Actions No. 3030, Cambridge, UK, 1988.

[BUR 97] BURNARD L., BAKER P., MCENERY A., WILSON A., An analytic framework for the validation of language corpora, Report of the ELRA Corpus Validation Group, Paris, 1997.

[CAL 91] CALZOLARI N., "Lexical databases and textual corpora: perspectives of integration for a Lexical Knowledge Base", in ZERNIK U. (ed.), *Lexical Acquisition: Using on-Line Resources to Build a Lexicon*, Erlbaum Ass., New York, 1991.

[CAL 01] CALZOLARI N., LENCI A., ZAMPOLLI A., BEL N., VILLEGAS V., THURMAIR G., "The ISLE in the Ocean. Transatlantic standards for Multilingual Lexicons (with an eye to machine translation)", *Proceedings of MT Summit VIII*, Santiago De Compostela, Spain, 2001.

[CAL 96] CALZOLARI N., MC NAUGHT J., ZAMPOLLI A., Eagles Final Report: EAGLES Editors' Introduction, EAG-EB-EI, Pisa, 1996.

[CAL 99] CALZOLARI N., ZAMPOLLI A., "Harmonised large-scale syntactic/semantic lexicons: a European multilingual infrastructure", *MT Summit Proceedings*, Singapore, pp. 358–365, 1999.

[CAL 02] CALZOLARI N., ZAMPOLLI A., LENCI A., "Towards a standard for a multilingual lexical entry: the EAGLES/ISLE initiative", in GELBUKH A.F. (ed.), *Computational Linguistics and Intelligent Text Processing, 3rd International Conference, CICLing 2002*, Mexico City, Mexico, Springer, pp. 264–279, 17–23 February, 2002.

[EAG 96a] EAGLES, Evaluation of natural language processing systems, Final Report, Center for Sprogteknologi, Copenhagen, 1996.

[EAG 96b] EAGLES Subcategorization Standards, EAGLES, CNR-ILC, Pisa, 1996.

[EAG 99] EAGLES Recommendations on Semantic Encoding, EAGLES, CNR-ILC, Pisa, 1999.

[FRA 06] FRANCOPOULO G., GEORGE M., CALZOLARI N., MONACHINI M., BEL N., PET M., SORIA C., "Lexical markup framework (LMF)", *Proceedings of LREC 2006*, Genova, Italy, ELRA, pp. 233–236, 2006.

[GEN 94] GENELEX, Report on the Semantic Layer, Project EUREKA GENELEX, Version 2.1, 1994.

[GIB 97] GIBBON D., MOORE R., WINSKI R., *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, New York, 1997.

[HEI 91] HEID U., MCNAUGHT J., EUROTRA-7 study: feasibility and project definition study on the reusability of lexical and terminological resources in computerised applications, Final report, 1991.

[KHA 93] KHATCHADOURIAN H., MODIANO N., "Use and importance of standard in electronic dictionaries: the compilation approach for lexical resources", *Literary and Linguistic Computing*, vol. 98, Oxford University Press, 1993.

[LEE 96] LEECH G., WILSON A., Recommendations for the morphosyntactic annotation of corpora, EAG-TCWG-MAC/R, Lancaster, 1996.

[LEN 99] LENCI A., BUSA F., RUIMY N., GOLA E., MONACHINI M., CALZOLARI N., ZAMPOLLI A., Linguistic specifications, SIMPLE Deliverable D2.1., CNR-ILC and University of Pisa, 1999.

[MON 96] MONACHINI M., CALZOLARI N., *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages*, EAGLES, CNR-ILC, Pisa, 1996.

[RUI 98] RUIMY N., CORAZZARI O., GOLA E., SPANU A., CALZOLARI N., ZAMPOLLI A., "The European LE-PAROLE project: the Italian syntactic Lexicon", *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spain, ELRA, pp. 241–248, 1998.

[TAU 11] TAUS, Report on a TAUS research about translation interoperability, 25 February, 2011.

[UND 97] UNDERWOOD N., NAVARRETTA C., A draft manual for the validation of Lexica, Final ELRA Report, Copenhagen, 1997.

[WAL 87] WALKER D., ZAMPOLLI A., CALZOLARI N. (eds), Towards a polytheoretical lexical data base, CNR-ILC Report, Pisa, 1987.

[WAL 95] WALKER D., ZAMPOLLI A., CALZOLARI N. (eds), *Automating the Lexicon*: *Research and Practice in a Multilingual Environment*, Oxford University Press, Oxford, 1995.