Harmonised large-scale syntactic/semantic lexicons: a European multilingual infrastructure

Nicoletta Calzolari
Istituto di Linguistica Computazionale
(ILC) del CNR, Pisa, Italy

Antonio Zampolli
Istituto di Linguistica Computazionale
(ILC) del CNR, Pisa, Italy

Abstract

The paper aims at providing an overview of the situation of Language Resources (LR) in Europe, in particular as emerging from a few European projects regarding the construction of large-scale harmonised resources to be used for many applicative purpose, also of multilingual nature. An important research aspect of the projects is given by the very fact that the large enterprise described is, at our knowledge, the first attempt at developing wide-coverage lexicons for so many languages (12 European languages), with a harmonised common model, and with encoding of structured "semantic types" and semantic (subcategorisation) frames on a large scale. Reaching a common agreed model grounded on sound theoretical approaches within a very large consortium is in itself a challenging task. The actual lexicons will then provide a framework for testing and evaluating the maturity of the current state-of-the-art in lexical semantics grounded on, and connected to. a syntactic foundation. Another research aspect is provided by the recognition of the necessity of accompanying these "static" lexicons with dynamic means of acquiring lexical information from large corpora. This is one of the challenging research aspects - for the next years - of a global strategy for building a large and useful multilingual LR infrastructure.

1 Introduction

The paper aims at providing an overview of the situation of Language Resources (LR) in Europe, in particular as emerging from a few representative European projects regarding the construction and acquisition of large-scale harmonised resources to be used for many applicative purpose, also of multilingual nature. The structure of the paper is as follows: we provide here in the introduction some historical notes, in section 2 we describe two large projects concerning the building of wide coverage syntactic and semantic lexicons: section 3 introduces the distinction between static and dynamic lexicons, which is then exemplified in section 4 with the description of the project SPARKLE; section 5 gives a few conclusive remarks on the European standardised multilingual infrastructure for LR.

1.1 Some historical background

"The tendency predominant in the 70's and in the first half of the 80's, to test linguistic hypotheses with small amounts of (allegedly) critical data, rather than to study extensively the variety of linguistic phenomena occurring in communicative contexts" (Godfrey, Zampolli 1997) has certainly contributed to the scarcity of interest, in the same period, of the Natural Language Processing (NLP) sector for the creation and the analysis of large corpora and the construction of extended lexicons.

The 1986 Grosseto (Tuscany) Workshop "On automating the lexicon" (Walker *et al.* 1995) is usually recognised as the event marking an inversion of tendency and the starting point of the process which gradually brought the major actors of the NLP sector to pay more and more attention to reusable language resources. This process, which was fostered by a number of initiatives which followed directly from the Grosseto workshop, achieved a crucial step through the recognition, in the so-called Danzin Report (1992), of the infrastructural role of LR (see also Zampolli 1991). This was very influential in the formation of the strategy of the European Commission (EC). In fact, the issue of LR is now regularly present in the initiatives of the EC in the field of language processing.

As other current actions in the field of LR (e.g. EAGLES, which is a direct descendant of the "Pisa Group", set-up at the Grosseto Workshop by the Istituto di Linguistica Computazionale (ILC) of Pisa and sponsored by the Association for Computational Linguistics

(ACL) to explore the feasibility of "polytheoretical lexicons" (Walker *et al.* 1987)), the PAROLE and SIMPLE projects, building large corpora and lexicons for many European languages, are the follow-up of some initiatives promoted at the Grosseto Workshop. The Council of Europe, which had co-sponsored the workshop, formed a group of experts, representing European institutes with a well established tradition in the field of lexical and corpus studies, to explore the feasibility of harmonising their activities, in order to establish a Network of European Reference Corpora (NERC, for which see Calzolari, Baker, Kruyt 1996; Zampolli 1996). This group, gradually enlarged to include members of all the European Union (EU) languages, constituted the PAROLE Consortium which has executed the LE (Language Engineering) PAROLE project now followed by the LE SIMPLE project carried on by a similar Consortium[1].

2 Goals of PAROLE and SIMPLE

2.1 PAROLE: Corpora and Morphological and Syntactic Lexicons

The central goal of LE PAROLE was to produce in Europe an initial core of harmonised corpora and lexicons. For each of the following languages a corpus of at least 20 million words and a lexicon of 20.000 entries was produced: Catalan, Danish, Dutch, English, Finnish, French. German, Greek. Italian, Portuguese, Spanish (lexicon only), Swedish. In addition, a corpus of respectively 20, 15, 3 million words was produced for Belgian-French, Irish, Norwegian.

The main characteristics of the corpora and lexicons can be summarised as follows:

[1] PAROLE and SIMPLE are projects sponsored by EC DGXIII in the framework of the Language Engineering program. The PAROLE and SIMPLE Consortia are formed by the following partners: Erli (now Lexiquest)-Paris, Institute for Language and Speech Processing-Athens, Institut d'Estudis Catalans, University of Birmingham, Univ. of Sheffield, Det Danske Sprog- og Litteraturselskab, Center for Sprogteknologi-Copenhagen, Språkdata-Göteborgs Universitet, University of Helsinki, Instituut voor Nederlandse Lexicologie-Leiden, Université de Liège BELTEXT, Centro de Linguistica da Universidade de Lisboa, Instituto de Engenharia de Sistemas e Computadores-Lisboa, Fundacion Bosch Gimpera Universitat de Barcelona, Institut für Deutsche Sprache; in addition Consorzio Pisa Ricerche (coordinator of PAROLE), Institiúid Teangeolaiochta Éireann-Dublin and Institut National de la Langue Française-CNRS were in PAROLE, while Università di Pisa (coordinator of SIMPLE). Istituto di Linguistica Computazionale di Pisa, and University of Graz are in SIMPLE.

Corpora

Encoding: All the information explicitly represented in the source texts is encoded following essentially the CES (Corpus Encoding Standard) designed by EAGLES, on the basis of the Text Encoding Initiative (TEI) guidelines (Ide et al. 1996). 250,000 running words are tagged at the morphosyntactic level, following the EAGLES guidelines (Leech, Wilson 1996: Monachini, Calzolari 1996), instantiated by each PAROLE partner for his own language.

Common model: The compatibility and interchangeability of the various corpora is ensured by the adoption of commonly defined criteria for *i)* corpus design, *ii)* composition as regards text types and their percentage in the corpus, *iii)* encoding and *iv)* linguistic annotation.

Availability: Each corpus is accessible for consultation, possibly via INTERNET. A subset of 3 million words of each corpus (including the tagged words) are also distributable through ELRA: i.e., a physical copy of it can be given to the users. Restrictions on the type of uses depend on the restrictions imposed by the holders of the copyright of the source texts, when they have authorised the inclusion of their texts in the corpus.

Lexicons

Model: The PAROLE Lexicon model for Morphosyntax and Syntax is based on the results of EAGLES (Sanfilippo *et al.* 1996) and EUREKA GENELEX (Antoni-Lay *et al.* 1994), further developed within the PAROLE project (Calzolari, Montemagni, Pirrelli 1996). Thanks to that all the lexical resources are declarative, theory and application independent, harmonised, multifunctional, and able to evolve easily, for example to incorporate other levels of information or to become multilingual. This approach, which answers to the requisites of generality, explicitness, and variability of granularity, guarantees a large scale reusability.

The model - given the high level of precision in the description - is in fact designed to ensure that application dependent models of data and applicative dictionaries can be derived from this general repository of information. An application model - with its own specificities - can be derived from the generic one through 'mappers' which are part of the lexicon tool.

The coverage is 20,000 entries per language fully described at the morphological and syntactic levels.

The current PAROLE lexical resources encode the following morphological and syntactic information:

- Morphology:
- written forms (graphical unit) including stems and
- variants
- morphosyntactic category (part of speech) and as appropriate a sub-category
- inflected forms
- morphological features
- derivation
- abridged forms
- Syntax:
- subcategorisation patterns (with optionality)
- grammatical relations of subcategorised complements
- control
- diathesis and lexical alternations
- pronominalisation
- linear order constraints
- constraints on the syntactic context where the lexical entry is inserted
- syntactic compounds (idioms, etc.)

The information encoded is divided into optional and mandatory classes for entries. Not all languages need to use all the optional information, while the mandatory one is required for all (the categories in the DTD are the union of all the categories required for the different languages).

Format and Tools: The exchange format for the lexicons • as for the corpora - is SGML: all the lexicons share the same DTD for the morphological and syntactic layers. Moreover, the use of a common set of lexicon management tools is a guarantee that all lexicons fully conform to the model. The use of these tools was a precondition of an industrial level of quality for the volumes of data (in so many languages) that PAROLE has delivered.

Availability: All the lexicons are publicly available, through ELRA.

2.2 SIMPLE: Semantic Lexicons

Semantics is today - and will be in the next years - the crucial and critical issue in Human Language Technology (HLT). Every application having to manage with information, in the ever growing importance of the so-called 'content industry', calls for systems which go beyond the syntactic level to understand the 'meaning'. Many theoretical approaches are tackling different aspects of semantics, but in general they still have to be tested *i)* with really large-size implementations, and *ii)* with respect to their actual usefulness and usability in real-world systems both of mono- and multi-lingual nature. SIMPLE arms at addressing directly point *i)* above, while providing the necessary platform to allow future projects to address point *ii)*.

Even more when we consider the multilingual aspect, with its problems and challenges - which is today the focus of attention in HLT programs in Europe and worldwide - again semantics is at stake. We cannot hope to successfully address the multilingual aspect without some solution to the semantic aspect (unless we use only statistical techniques). For the addition of a multilingual layer (multilingual links) to available lexical resources it is essential to have a 'harmonised' set of semantic lexicons addressing in a uniform way the core of what is needed for NLP, i.e. "semantic typing" of heads and arguments, which is at the centre of the SIMPLE project.

SIMPLE positions itself inside the strategic policy - supported by the EC - which aims at providing a core set of language resources for the EU languages. The SIMPLE project, a follow up to PAROLE, aims in fact at adding a semantic layer to a subset of the existing morphological and syntactic layers. The semantic lexicons (covering about 10,000 word meanings) are being built in a harmonised way for all the 12 languages covered by PAROLE (see above). The main types of information to be encoded for nouns, verbs, and adjectives are: domain information, the semantic type of the head (with a structured template-type), and the semantic type of the arguments of predicates (to be defined at different levels of granularity) which are linked to the syntactic arguments.

The semantic lexicons will be partially corpus-based, exploiting the harmonised and representative corpora built within PAROLE. This will make the semantic encoding aware of actual corpus distinctions and not of potential, and often misleading, abstract generalisations based on linguist/lexicographer's introspection only.

The SIMPLE project represents - at our knowledge - the first attempt to tackle harmonised encoding of semantic types and semantic (subcategorisation) frames on a large scale, i.e. for so many languages and with wide coverage. Even though it is a real lexicon building project, it must be seen as addressing challenging research aspects and will provide a framework for testing and evaluating the maturity of the current state-of-the-art in the realm of lexical semantics grounded on, and connected to, a syntactic foundation.

The SIMPLE lexicons, even though monolingual, are designed having in mind their future cross-lingual linking. All the lexicons do in fact share and are built using the same core ontology and the same set of semantic templates. These lexicons will be the essential basis for any future European multilingual initiative in the lexical area aimed at NLP and LE applications.

In the specification phase we have taken into account requirements of NLP applications and tasks (parsing, generation, machine translation, word sense disambiguation, cross-language information retrieval, etc.) - also as stated in the EAGLES report of the Lexicon/Semantics Working Group (Sanfilippo *et al.* 1999) - for the decisions on the basic semantic notions and the more specific types of semantic information to be encoded. This is of utmost importance given the applicative objectives of the PAROLE/SIMPLE lexicons. A dichotomy at stake here is the one between generality of a LR vs. usefulness for applications. In principle, only when we know the actual specific use we intend to do of a LR can we build the 'very best' LR for that use, but this has proved to be too expensive and not realistic. In practice, however, there exists a large core of information that can be shared by many applicative uses, and this leads to the concept of "generic" LR, which is at the basis of the EAGLES initiative and of the PAROLE/SIMPLE projects. This generic shareable core of information must then be enhanced and tuned with other means (see sections 3 and 4).

The model: basic issues

In the specification phase of the project the formal representation of the "conceptual core" of the lexicons was designed, and the basic structured set of "meaning-types" - i.e. the core ontology - to be used as a common starting point and a shared device to build the harmonised language specific semantic lexicons was defined (see Busa *et al.* 1999). Such a task has tackled questions that are at the core of lexical semantics research. The development of twelve harmonised semantic lexicons requires strong mechanisms for guaranteeing uniformity and consistency of the representations. These mechanisms, in turn, guarantee that within the same language consistent formal devices apply cross-domain and cross-categorially. Finally, the multilingual component translates into the requirement of identifying elements of the semantic vocabulary for structuring word meaning which are at the same time independent from any individual language but able to capture linguistically useful generalisations for different NLP tasks.

A coherent development of semantic lexical resources must be guided by an underlying theoretical framework for structuring word meaning and generating concepts which satisfies both ontological considerations as well as the need to capture linguistic generalisations. The SIMPLE model is a concrete major step towards this objective. It is based on EAGLES Lexicon/Semantics Working Group recommendations (Sanfilippo *et al.* 1999) and on extensions of Generative Lexicon (GL) theory (Pustejovsky 1998). An essential characteristics - which makes it basically different from EuroWordNet (where the main structuring semantic relations are synonymy and hyponymy) - is its ability to capture the various dimensions of word meaning which are equally important in language and therefore in the development of a computational lexicon. The basic vocabulary relies on an extension of "qualia structure" for structuring the semantic/conceptual types, which is understood as a representational tool for expressing the componential multidimensional aspect of word meaning (Pustejovsky 1991, 1995; Calzolari 1991).

The perspective adopted is that all words have internal structure, based on different semantic types, and differ in terms of complexity, which affects the way they compose in a sentence. The so-called "extended qualia structure" of SIMPLE addresses the concern of capturing more or less subtle linguistic differences while maintaining a systematic and consistent structuring of the lexical representations. This is achieved by specifying, for each qualia role, its extended *qualia set,* namely subtypes of that role which are consistent with its interpretation. To the standard approach of defining semantic classes along one dimension - which fails at capturing underlying generalisations along different dimensions -, we have thus opposed a framework whose development has been crucially concerned with capturing the multidimensionality of meaning. Assuming that lexical items differ according to which dimension of meaning carries most of the semantic weight, the GL-SIMPLE model also clarifies the nature of the underspecification of certain items, which may be highly underspecified along one dimension while providing a rich semantic contribution along other dimensions.

For purposes of combining the ontology and the theoretical framework with the practical lexicographic task of encoding the lexicon, we have created a "library" of template-types, which reflect the well-formedness conditions of a given type and provide the constraints for a lexical item belonging to that type. In the encoding of the lexicon for a given language, the linguists/lexicographers have available the common set of language independent template-types providing the "blueprint" of any given type. The relevance of this approach for building consistent resources is that types both provide the formal specifications and guide subsequent encoding, thus satisfying theoretical and practical methodological requirements.

The SIMPLE model thus allows to consistently generate concepts out of a set of ontological categories that are grounded in linguistic behaviour. The model has a high degree of generality in that it provides the same mechanisms for generating broad-coverage and coherent concepts independently of their grammatical/semantic category (entities, events, qualities, etc.), an aspect which is often lacking in existing lexicons, where the focus is of-

ten on the representation of the clear, well-known cases while the semantics of more complex cases is neglected.

Cross-lingual uniformity

The main criterion for the selection of the senses to be encoded in SIMPLE is the frequency of occurrence in available PAROLE text corpora. The fact that these corpora share a common design with respect to text types and *genres* for all the languages ensures some uniformity in vocabulary (sense) selection.

In order to achieve overlapping of a subset of the senses for all the 12 languages - very important for future multilingual linking - it was decided to reuse the so called "base concepts" of EuroWordNet (after some 'cleaning') as a common set of senses to be encoded for all the languages. This set of rather generic senses (i.e. of high level in the taxonomy) constitutes the common core from which started the encoding phase, and to which more specific senses extracted from text corpora are linked. For these senses a cross-lingual link for all the 12 languages is already automatically given through their link to the English EuroWordNet Interlingual Index. This set, from which all the other senses depend, also guarantees uniformity of coverage in extension - i.e. with respect to different semantic classes - for all the languages, and will allow easier cross-language comparison and evaluation of encoding among languages during the project.

3. "Static" lexicons vs. "dynamic" acquisition of lexical information

The resources built in these projects do provide the essential basic infrastructure, but it is well recognised that they do not have enough coverage, not only for practical reasons, but for more structural and inherent reasons. No "static" resource can ever be adequate and satisfying, from more than one perspective: *i)* in extension: it cannot obviously cover new formations, or all the possible domains, and *ii)* in depth: not even for the existing lexical entries it can provide all the necessary and useful linguistic information (e.g. not necessarily all the subcategorisation types actually occurring in a specific domain are covered by a general lexicon). For them to become really usable, it is essential that these generic, core LR are built in such a way that *i)* they are really open to different types of enrichments and customisations, possibly to be done in an automatic way, and *ii)* the information is granular enough so that different applications can extract what they need in the format they need.

The common generic platform of LR - constituting the basic infrastructure - needs therefore, for the reasons said above, to be enhanced and fine-tuned in various ways (according to the domain, to the task, to the system (Information Retrieval, Machine Translation, ...), etc.) to become actually usable within specific applications. This makes it vital, for any sound lexicon development strategy, to accompany core static lexicons with dynamic means for enriching and integrating them - possibly on the fly - with the types of information which are known to be structurally and intrinsically missing from existing available LR. This global view eliminates another apparent dichotomy, i.e. the one between static vs. dynamically built (or incremental) resources, encompassing the two approaches in a more comprehensive perspective that sees the two as complementary and equally necessary facets of the same problem.

Steps towards this objective have been taken - in the past few years - by a consistent number of groups all over the world, with many varied research and development efforts aimed at acquiring linguistic and, more specifically, lexical information from corpora. The next section describes the approach towards a dynamic lexicon taken within the SPARKLE project.

4. The research project SPARKLE

Among the EC projects working in this direction we mention LE SPARKLE (Shallow PARsing and Knowledge extraction for Language Engineering[2]), combining shallow parsing and lexical acquisition techniques capable of learning (from large corpora) aspects of word knowledge required for LE applications (Federici *et al.* 1998). The project (http://www.ilc.pi.cnr.it/sparkle.html) is positioned as research on the development of methodologies and techniques for application- or domain-dependent lexical resources to be acquired (semi)-automatically from texts, an area which is crucial to most NLP applications. Economically feasible development of language models and of substantial lexical resources for real-world NLP applications needs to be based on substantially (semi)-automated techniques and flexible tools for analysing and extracting lexical information from textual corpora, otherwise coverage and/or accuracy will remain inadequate.

Computational lexicons, like human dictionaries, often represent a sort of stereotypical/theoretical language. Carefully constructed and selected large corpora are essential sources of linguistic knowledge for the extensive description of the concrete use of the language in real texts: this is impossible relying on introspection only and on native speakers' (even if lexicographers) intuition.

[2] Also SPARKLE is a LE project sponsored by EC DGXIII. Its partners are: Consorzio Pisa Ricerche (coordinator), Universities of Cambridge, Sussex, and Stuttgart, Xerox Research Centre Europe, Sharp Laboratories of Europe, Daimler-Benz.

Evidence of actual usage is in fact frequently in contrast with what one would expect based on introspection only. To be habitable and practical, a computational lexicon has to faithfully represent the apparently 'irregular' facts (evidenced by corpus analysis), and the divergences of actual usage from what is potentially/in theory acceptable. We need to clearly represent - and separate - what is allowed, but only very rarely instantiated, with respect to what is both allowed and actually used. With respect to this issue, a number of dichotomies have to be considered not as opposite views, but as complementary perspectives:

- rules vs. tendencies
- absolute constraints vs. preferences
- discreteness vs. continuum/gradedness
- theoretical/potential vs. actual
- intuition/introspection vs. empirical evidence
- theory-driven vs. data-driven
- paradigmatic vs. syntagmatic
- symbolic vs. statistical.

We claim that the second element of the above dichotomies has to be highlighted, in order then to combine the two. To this end. more robust and flexible tools are needed for (semi-)automatic induction of linguistic knowledge from texts. This usually implies a bootstrapping methodology, because extraction presupposes some capability of automatically analysing the raw text in various ways, which first requires a (partial) lexicon. A cyclical methodology is of help in getting out from this loop. The induction phase must however be followed by a linguistic analysis and classification phase, if the induced data is to be used and merged together with already available resources, so that it can contribute to enrich them.

The SPARKLE project has shown (see Briscoe *et al.* 1999) how far simple robust phrasal parsing combined with classification techniques utilising limited and manageable linguistic knowledge and statistical data from substantial corpora can ameliorate this problem in the area of predicate subcategorisation, argument structure and semantic preferences; an area in which most extant conventional dictionaries, lexical databases and realistic lexicons are demonstrably weak or - when available - by necessity never complete.

To satisfy this requirement, SPARKLE has followed two parallel tracks of development: *i)* shallow parsers have been developed to produce a phrasal-level syntactic analysis of naturally occurring free text in the 4 languages of the project (French, English, German, and Italian), and *ii)* lexical acquisition systems have been built (based on statistical and analogy-based techniques) capable of learning - in a (semi-)automatic way - those aspects of word-knowledge, derived from free text, which are needed for LE applications and are not found - or at least not exhaustively - in conventional dictionaries.

Common annotation schemes have been defined for three levels of syntactic analysis (chunking, phrasal parsing, and functional annotation) for all the languages, and a common description language for lexical encoding has been designed. These technical standards are at the basis of a common evaluation framework defined both for the parsers and for the lexical acquisition systems.

The background applicative rationale for a project such as SPARKLE can be sought in the ever growing necessity within the Multilingual Information Society of providing accurate and immediate access, consumption, exchange and dissemination of multilingual information, accessible through telematic systems and services. SPARKLE'S main applicative objective was to address these requirements by developing robust and portable tools leading to commercial applications devoted to the management of multilingual information in electronic form. The parsers and lexicons produced have been used in fact by the industrial partners to build pilot applications in the areas of cross-lingual information retrieval and speech dialogue recognition.

5. A European standardised multilingual LR infrastructure

The availability of these large, uniformly structured lexical resources in so many EU languages - accompanied by means of enriching them - will offer the users the benefits of a standardised base. According to the subsidiarity concept, which is at the basis of many EU initiatives, the process started at the EU level is continued at the national level. This is already happening for a number of languages. The PAROLE/SIMPLE Lexicons and Corpora are being enlarged and extended to real-size lexicons at the national level in the framework of a number of National Projects for at least the following languages: Catalan, Danish, Dutch, Greek, Italian, Portuguese, Spanish, Swedish. These national initiatives show that the goal of the LR EC projects, aiming at providing a core set of resources to be extended with national support, is perfectly satisfied.

The fact that all these LR will be based on the existing models and standards defined and implemented at the European level will create a really large infrastructure of harmonised LR throughout all Europe. This achievement is of major importance in a multilingual country like Europe, where all the difficulties connected with the task of LR building are multiplied by the language factor. This would have been absolutely impossible without the fundamental role played by the EC LR and standards projects.

The thesis of this paper is the recognition of the essential infrastructural role that LR play in Human Language Technology, as the necessary common platform on which new technologies and applications must be based. In order 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 standardised wide-coverage LR is even stronger. Another tenet is the recognition of the need of a global strategic vision, encompassing different types of (and methodologies of building) LR. for an articulated and intelligent development of this field.

References

Antoni-Lay, M.H., Francopoulo, G., Zaysser L. (1994). "A Generic Model for Reusable Lexicons: The Genelex Project". In *Literary and Linguistic Computing,* 9 (1) (47-54).

Briscoe, T., Korhonen, A., Calzolari, N., Montemagni, S., Pirrelli, V., Federici, S., Carroll, G., Light, M., McCarthy, D., Prescher, D., Riezler, S., Rooth, M. (1999). *Syntactic and Semantic Type and Selection.* SPARKLE Deliverable 5.2.

Busa, F., Calzolari, N., Lenci, A., Pustejovsky, J. (1999). "Building a semantic lexicon: structuring and generating concepts". In *Proceedings of the 3rd International Workshop on Computational Semantics.* Tilburg.

Calzolari, N. (1991). "Acquiring and Representing Semantic Information in a Lexical Knowledge Base". In J. Pustejovsky (ed.). *Proceedings of the Workshop on Lexical Semantics and Knowledge Representation.* Berkeley, CA.

Calzolari, N. (1998). "An overview of Written Language Resources in Europe: a few reflections, facts, and a vision". In *Proceedings of the First International Conference on Language Resources and Evaluation.* Granada. (217-224).

Calzolari, N., Baker, M., Kruyt, T. (eds.). (1996). *Towards a Network of European Reference Corpora: Report of the NERC Consortium Feasibility Study. Linguistica Computazionale XI.* Giardini, Pisa.

Calzolari N., Montemagni S., Pirrelli V. (1996). *Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon: Verb Subcategorisation.* MLAP PAROLE, Pisa.

Danzin, A. (1992). Groupe de réflexion stratégique pour la Commission des Communautés Européennes (DG XIII), *Vers une infrastructure linguistique européenne.* Document available from DG XIII-E, Luxembourg.

Federici S., Montemagni S., Pirrelli V., Calzolari N. (1998). "Analogy-based Extraction of Lexical Knowledge from Corpora: The SPARKLE Experience". In *Proceedings of the First International Conference on Language Resources and Evaluation.* Granada. (74-82).

Godfrey, J., Zampolli, A. (1997). "Resources - Overview", Chapter 12. In Varile, Zampolli (eds.).

Ide, N., Veronis, J., Priest-Dorman, G. (1996). *Corpus Encoding Standard.* EAGLES/MULTEXT.

Leech, G., Wilson, A. (1996). *Morphosyntactic annotation.* EAGLES.

Monachini, M., Calzolari, N. (1996). *Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages.* EAGLES, Pisa.

Palmer, M. *et al.* (1999). *Multilingual Resources.* See http://www.cs.cmu.edu/~ref/mlim/

Pustejovsky, J. (1991). "The generative lexicon". In *Computational Linguistics.* 17 (4) (409-441).

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, MA, MIT Press.

Pustejovsky, J. (1998). "Specification of a Top Concept Lattice". Ms. Brandeis University.

Sanfilippo, A. *et al.* (1996). *EAGLES Subcategorization Standards.* See http://www.icl.pi.cnr.it/EAGLES96/syntax/syntax.html

Sanfilippo, A. *et al.* (1999). *EAGLES Recommendations on Semantic Encoding.* See http://www.ilc.pi.cnr.it/EAGLES96/rep2

Varile, G.B., Zampolli, A. (eds.) (1997). *Survey of the State of the Art in Human Language Technology.* Sponsored by the Commission of the European Union and the National Science Foundation of the USA. Giardini Editori. Pisa and Cambridge University Press.

Walker, D., Zampolli, A., Calzolari, N. (eds.) (1987). *Towards a Polytheoretical Lexical Data Base.* Pisa, Istituto di Linguistica Computazionale.

Walker, D., Zampolli, A., Calzolari N, (eds.) (1995). *Automating the Lexicon: Research and Practice in a Multilingual Environment.* Proceedings of a Workshop held in Grosseto. Oxford University Press, Oxford.

Zampolli, A. (1991). *Preliminary Considerations on the Constitution of an ELTA (European Language Technology Agency).* Pisa, Document prepared for DG XIII.

Zampolli, A. (1996). "Introduction". In Calzolari, Baker, Kruyt (eds.).