Nicoletta Calzolari, Istituto di Linguistica Computazionale - CNR - Pisa

Lexicon and Corpus: a Multi-faceted Interaction

1. Computational Lexicons for Language Engineering

All Language Engineering (LE) applications require knowledge about words. The first basic operation that any NLP system must perform consists in "recognizing" the words of the input-text. This means i) to search for the corresponding entry in a Computational Lexicon for each word (or multi-word) of a text, and ii) to associate the linguistic information provided by the lexical entry and relevant for that particular application, to the word in the text. Moreover, in order to be practical and to be capable of performing with some hope of success, LE systems must be furnished with a large-size lexicon, covering a realistic vocabulary, and providing the types of linguistic knowledge required for the application. A survey of the types of linguistic knowledge needed for different systems is found in the final report of the EC Eurotra-7 project (Heid, McNaught, 1991).

If, in addition, we do not want to build a new lexicon for each new application or system, we need to build large, generic and "reusable" lexicons (Calzolari, 1991), from which the required data – in the required format – can be extracted by different applications – through appropriate filtering and conversion procedures.

It is a matter of fact that, given the present state-of-the-art, the linguistic information required for real-life applications, which has to be encoded in a Computational Lexicon, possibly in a standardised or normalised way, can be very complex and difficult to acquire/gather, to structure, and to represent in a formal way. Even though many steps forward have been made in the last ten years as regards computational lexicons, we are still in a position to reiterate that the lexicon is a "major bottleneck" for natural language processing (NLP) systems. The capabilities of NLP systems were and have remained weak because of the labour intensive nature of encoding lexical entries.

After approximately ten years of acquiring (semi-)automatically lexical/linguistic information from machine-readable dictionaries (see Calzolari, Briscoe, 1995 for an overview of the ACQUILEX project which was started exactly from this hypothesis of work), we can clearly see not only the strong points but also the intrinsic limitations of this workplan. It is not possible to extract what is not present in current dictionaries, and dictionaries lack many types of crucial information, besides being incomplete, partially incoherent and sometimes unreliable for the information they contain. This is not an attempt to destroy the work done by many research groups – our own among the first (Calzolari, 1982) – in this field, but simply to recognise that, as useful as it was, and still is, it needs to be complemented by other types of lexicon building efforts. This direction of work was, and is, just one piece of the overall system needed if we want to aim at building a usable lexicon.

In addition to recognizing and acknowledging the partial insufficiency of the above method, at the same time it became apparent that it was more and more feasible to treat very large text corpora with (semi-)automatic methods. In the recent past, both (computational) lexicographers and NLP researchers have advocated the use of corpora for (semi-)automatic acquisition of lexical information.

2. Why to resort to written and spoken Corpora for building Computational Lexicons?

Carefully constructed, large written and spoken corpora are essential sources of linguistic knowledge if we hope to provide extensive and adequate descriptions of the concrete use of the language in real text. These types of descriptions certainly remain impossible if we only rely on introspection and native speaker intuition (see Calzolari, forthcoming, for many details on this). This is true for both of the main approaches within NLP and Speech systems in use today: the rule-based approach and the statistical approach.

Presently, corpora are recognized, by more and more research and development groups, as the most precious aid in designing systems that respond to user needs, in terms of types of texts and real language to be treated. We could say that this is the trend or even the fashion of today's language engineering systems, which seem destined to last for a while (see Zampolli 1995).

In the past both theoretical and computational linguists typically concentrated their efforts on evaluating particular properties of competing syntactic theories, studying peculiar linguistic phenomena, "interesting" for the comparison of the explanatory power of the models. However, the LE products needed in the current Multilingual Information Society require robust tools relying on robust components (lexica, grammars, etc.) based on an inventory and description of the variety of phenomena occurring in real texts, in the different communicative contexts in which a product will be used. Among such phenomena we must include e.g. structures that have been underestimated or underdiscussed in traditional main-stream linguistics (multi-word units, collocations, idioms, etc.), linguistically "uninteresting" phenomena occurring in real texts (dates, tables, titles, acronyms, etc.), "deviations" from the standard language described by linguistic models (repetitions, ellipses, abbreviated styles, "agrammaticalities", etc.). Corpora are full of these phenomena.

Furthermore, corpora are obviously the only source of data for acquiring statistical information, and for providing training data to construct stochastic models.

The study of corpora is also essential for the identification and characterization of sublanguages: when natural language is used in specific domains or communicative contexts, it may be restricted in the lexical, syntactic, semantic discourse properties, so that we can expect a significant contribution to the solution of ambiguities in such specific text types.

Quite recently, the essential role of corpora in the evaluation of models – also lexical models – and of NLP and Speech systems dependent on such models, has been highlighted. The use of corpora for system evaluation can help designers to develop better systems (in the self-organizing approach to train and extend the model, in the rule-based approach for the analysis of results). Test-bed corpora are now being constructed and/or used by almost every application project. Corpora can also be used to help the end-users in judging the comparative merits and demerits of systems they are interested in acquiring (see the final report of the EAGLES Evaluation Working Group, King et al., 1996).

Other Corpora uses that need to be mentioned are: language teaching and learning, literary, sociolinguistic, lexicographic, stylistic studies, etc.

3. Economical benefits vs. technical problems

Everything that was said above raises the need to at least attempt semi-automatic construction of a new generation of computational lexicons directly from corpora, otherwise coverage and/or accuracy will remain inadequate. However, in order for the development of new lexicons and/ or the amelioration of existing lexicons through the exploitation of the richness of linguistic information implicitly contained in real texts to be economically profitable, the exploitation needs to be based on substantially automated techniques for analysing and extracting lexical information from textual corpora. A number of fundamental problems in NLP will need to be solved before this highly desirable prospect becomes completely realistic and economically viable: the extraction of many types of information from corpora usually presupposes some partial capability to automatically analyse the raw text in various ways.

3.1 What kind of "regularities" ?

A classification method of the regularities, or "irregularities", that are evidenced through analysis of textual data is the correlation of different levels of linguistic analysis with different types of linguistic phenomena. It is evident that the correlation is not one-to-one, i.e. many phenomena are evidenced and adequately described through analysis at different levels of linguistic description. A partial display of the correlations between linguistic phenomena and levels of analysis is given in the figure below.

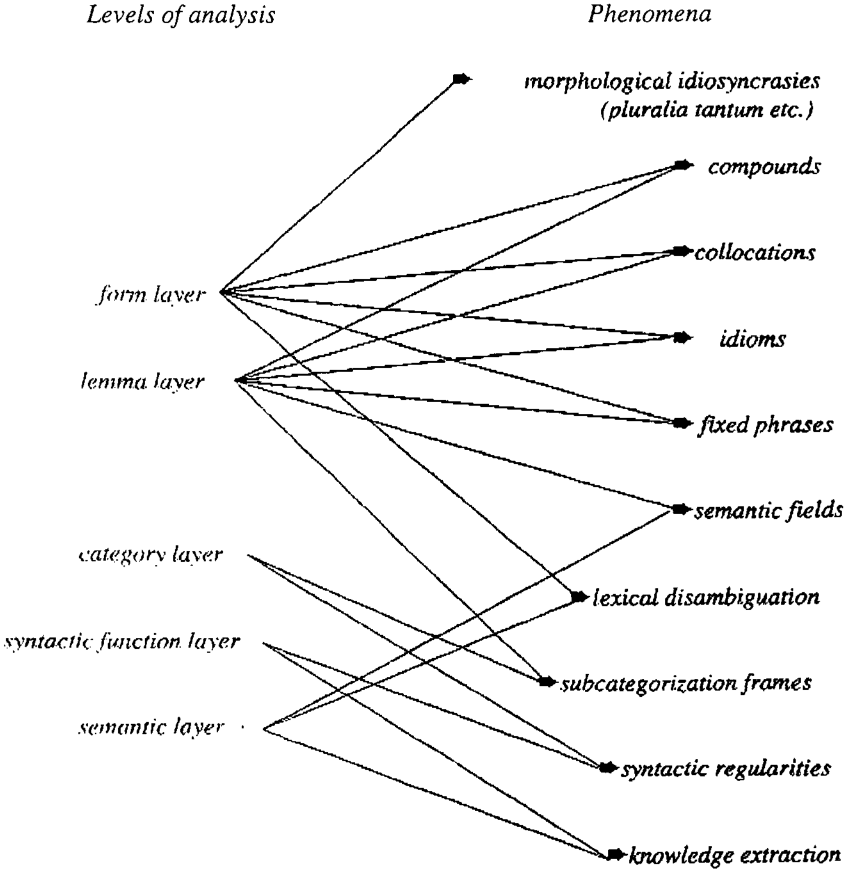

Figure 1. Correlations between linguistic phenomena and levels of analysis

3.2 Which techniques?

Given the state-of-the-art in our field, some of the linguistic analyses listed above can be performed automatically with good coverage and a good success rate (those at the top in the above figure), others allow at least semi-automatic processing, while the last ones (at the bottom) are more difficult to perform successfully either for coverage or for adequacy or both.

This means that *robust techniques* exist and are reliable in their performance for some types of corpus analysis, such as part-of-speech tagging or semi-automatic extraction and classification of many kinds of cooccurrences and collocations, others, e.g. phrasal parsing, will soon become reliable enough for massive use on large corpora of free text, and as work on statistical in particular and in general robust approaches to corpus analysis continues more complex analyses will become available for regular usage, e.g. lexical acquisition tools.

Already, the combined use of many of these techniques allows the extraction of information which complements that already available from MRD sources: an obvious example is the frequency of different types of linguistic phenomena at different levels.

The increasing availability and reliability of such techniques, and the ability to integrate them in opportune ways, will make the exploitation of text corpora of greater relevance in many LE tasks, from the acquisition of lexical information to the evaluation of models and systems.

4. Lexicon – Corpus Interactions

When we look attentively at the various ways in which lexicon and corpus are related to each other we cannot avoid highlighting the complexity of their mutual interactions. According to different perspectives, the relation goes in one of two possible directions, and in any case we cannot safely separate these two linguistic objects from one another as if they were independent entities. We can summarise, without claiming to be exhaustive, the lexicon (L) – corpus (C) interactions in the following list, where an arrow from L to C means, in general, the projection/ mapping of some lexical data on the corpus, while an arrow from C to L means acquisition of lexical information from corpora.

L → C	tagging	
C → L	frequencies	(of different linguistic "objects")
C → L	proper nouns	
L → C	parsing	
C → L	updating	
C → L	"collocational" data (MW, idioms, gram. patterns ...)	
C → L	"nuances" of meanings & semantic clustering	
C → L	lexical (syntactic/semantic) knowledge acquisition	
L → C	semantic tagging	
↓		
C → L	more semantic information on the lexical entry	
L → C	semantic disambiguation	
C → L	corpus based computational lexicography	
C → L	validation of lexical models	

Figure 2. Lexicon (L) - Corpus (C) interactions

5. Some experiences in EU projects: Issues and related problems

5.1. LRE EAGLES: Interdependence between Lexicon and Corpus Views

When we highlight the complex structure of the interrelationships between lexicon and corpus, we have to work on the assumption of an interdependence between the two views, and we have to take into account this interdependence in any lexical or corpus analysis or application.

This was, for example, the approach taken within the LRE EAGLES (see Calzolari, McNaught, 1996) project towards the development of standards both in Morphosyntax and Syntax: the awareness of the interdependence between lexical specifications and corpus tagsets / syntactic annotations has guided the formulation of the proposals in both the Corpus and the Lexicon Working Groups (see, for example, Monachini, Calzolari, 1996). Corpus tagging / annotating was considered as the first obvious application of a Computational Lexicon. Therefore attention was given to the definition of compatible sets of attributes and values (see also Heid, 1996).

5.2 LRE DELIS: Corpus-based Computational Lexicography

Within the LRE DELIS project, the design of the lexical entry was done with a combined approach: theoretical – the Fillmore frame semantics, and empirical – corpus evidence. In fact corpus data cannot be used in a simplistic way. In order to become usable they must be analysed according to some theoretical hypothesis, that would model and structure what would be otherwise an unstructured set of data. The best mixture of the empirical and theoretical approaches is the one in which the theoretical hypothesis is itself emerging from and is guided by successive analyses of the data, and is cyclically refined and adjusted to textual evidence.

The "frame semantics" approach defined by Fillmore (Fillmore and Atkins, 1992), was assumed as a theoretical modelling hypothesis, to be used both as a guide in the analysis of corpus data, and as a linguistic basis for the subsequent design of the lexical entry. Within this framework an essential descriptive strategy is defined as one which links together the semantic and syntactic descriptive levels. This is an essential characteristic of an approach to corpus analysis and to lexicon building which aims at reusing its results in the context of NLP applications.

Particular attention was paid to the correlation between different levels of linguistic description. In the project, the focus was on the correlation between syntactic and semantic aspects, but it was evident that other linguistic aspects – such as morphology, morphosyntax, lexical cooccurrence, collocational data, etc. – are closely interrelated, and these relations have to be captured when designing a lexical entry, and in particular when accounting for the phenomenon of meaning discrimination. It is the complexity of the interrelationships of all these aspects which makes semantic disambiguation such a hard task in NLP. One of our aims was to use textual corpora as a device to discover and reveal the intricacy of these relationships, and frame semantics as a device to unravel and disentangle the complex situation into elementary and computationally manageable pieces.

One of the most interesting – and intriguing – aspects of corpus use for a lexicographic task is the immediate evidence of the impossibility to use any type of description which is based on a clear-cut boundary between what is admitted and what is not. In actual usage of the language it is evident that its main characteristic is that of displaying a large number of properties which behave as a *continuum,* and not as properties of "yes/ no" type. The same is true for the so-called "rules", where we find more of a "tendency" towards a rule than a precise rule in corpus evidence. Most of the lexical (grammatical, syntactic, semantic) information must not be considered to be constraining information, but rather as preferential information. This creates problems at the level of the representation language, which must be able to accommodate this type of preferential information: this may not be easy and certainly not straightforward for a constraint-based formalism.

Another relevant aspect is the evidence of actual usage, frequently in contrast with what one would expect if one based his judgement only on introspection. A (computational or traditional) lexicon has to faithfully represent these "sometimes irregular" facts and these divergences of usage from what is potentially acceptable.

1) The first rule is that what is described in the lexicon cannot only be judged on the basis of native speaker's intuition, therefore leading to a description of a "theoretical language", rather than to the description of language as it is used.

2)The second rule is to allow – in the lexicon – for a clear representation of (and separation between) what is allowed, but only very rarely instantiated, and what is both allowed and actually used.

From corpus evidence, we get the impression of not having any clear-cut boundary in the analysis of many phenomena, but of language behaving more as a *continuum.* The same impression if we look at examples of different types of diathesis alternations: there are no clear-cut classes across languages or within one language (see Montemagni, Pirrelli, 1995). Again, it is more a tendency towards a rule than a precise rule.

Considering the most relevant phenomena that have emerged from the analysis of corpus evidence, the representation in the TFS lexicon seems to raise some basic issues and problems: the determination of the appropriate level of abstraction within the type hierarchy for each information type, the definition and representation of all the possible interactions between different kinds of information, the encoding of information that the current versions of HPSG usually do not deal with: semantic information, collocations, preferences, prototypicality, constraints, statistical information, etc.

5.3 ESPRIT BRA ACQUILEX: Extraction from texts vs. formal representation in lexicons

Also in the ACQUILEX project, where the text to be analysed for acquiring semantic and syntactic information was the text of natural language definitions in printed dictionaries, we encountered the problem of the mismatch between the vagueness and semantic density of natural language and the explicitness and poorness of the formal representation language. The rigour and lack of flexibility of a typed feature structure (TFS) representation language can cause difficulties when mapping it into natural language words – in particular word-meanings – ambiguous and flexible by their own nature. It is difficult to constrain word meanings within a rigorously defined organization: by their very nature they tend to evade any strict boundary.

Part of the results of meaning extraction become unmanageable at the formal level of the Type System. Many meaning distinctions, which can be generalised over lexicographic definitions and automatically captured, must be blurred into unique features and values (see Calzolari et al., 1993).

5.4 ET10 on COBUILD: Constraints or Preferences?

The same inadequacy of the formal machinery of a TFS representation language with respect to the complexity of the lexical information to be encoded emerged in the ET10 project on extracting syntactic/semantic information from the COBUILD dictionary. The necessity to formally represent all the information from COBUILD raises the problem of the distinction between constraining and preferential information. This distinction is not inherent in the nature of the data, but related to their use: the same grammatical specifications (e.g. number or voice) must be seen and used either as constraints or as preferences in different situations.

Preferential information is connected more to specific attributes: among others, usage indices specifying the register, style, variant; the type of action expressed by the verb; sortal semantic restrictions on complements and adjuncts, and on adjective collocates. All this information – but not only this – should not be treated as absolute constraints, whose violation makes a sentence totally unacceptable, but rather as preferences, making a given sentence more or less acceptable in a given context without affecting its grammaticality.

Unfortunately, despite some proposals in this direction, unification (or constraint)-based formalisms as they appear today do not easily capture the distinction (preferences are either ignored or treated as absolute constraints). Since weighted TFSs (weights associated with disjunctions of constraints) are not a viable solution for the time being, and since it is not the case of restricting the possible value(s) of an attribute to the most typical ones (with the result of excluding the less likely but still possible), an ad-hoc solution was used in the project (see Calzolari et al., 1995).

A method of tackling this problem could be the use of an analogy-based approach (see Federici and Pirrelli, 1994).

5.5 Analogy-based approach

This presents the advantages that both constraints and preferences are treated as clusters of nodes/features which are activated and operate according to the same principles. Therefore, the difference between constraints and preferences is not one of kind but of gradation. What differs is their range of application: a core pattern which is never unconfirmed is a constraint; a core pattern which can be unconfirmed in some cases is a preference. Actual extracted core patterns (from corpora or dictionaries) often combine constraints and preferences which can both be of a different nature. Preferential cores which are never unconfirmed when simultaneously activated can be considered as forming a constraint in its own right ("two preferences can make a constraint").

5.6 LE SPARKLE: Semi-automatic lexical acquisition from corpora

LE-Sparkle, a newly begun project, will address the problem of lexical acquisition from corpora by developing software which (semi-)automatically acquires lexical information. The central idea is to take advantage of the fact that text corpora contain hundreds or thousands of examples of word usages of intermediate frequency; by application of a partial parser, these examples are put in a form from which lexical information can be abstracted (see SPARKLE Technical Annex).

Technology for shallow parsing of naturally-occurring text – e.g. newspapers, technical documentation, and text accessible on the Internet – represents an attainable next step in the practical development of LE technology.

Building on the results of parsers, i.e. on a simple phrasal-level syntactic annotation of the texts, we aim at developing a lexical acquisition system capable of learning subcategorization, argument structure, argument/adjunct classification, semantic selectional preferences for individual predicates, diathesis alternation, from free text.

Combining many different types of state-of-the-art resources, tools, techniques (such as taggers, taxonomies of disambiguated nouns, normalized top levels of the taxonomies, robust phrasal parsers for real text corpora, statistical and analogy-based techniques), and using different approaches: statistical (e.g. Resnik, 1993), analogy based (e.g. Federici, Pirrelli, 1994), rule-based, for different tasks (such as sense disambiguation, semantic clustering, semantic acquisition of information, subject / object disambiguation) we think that we can make a step further in corpus analysis and in automatic acquisition.

For example, we want to use taxonomy nodes as "semantic tags" on text corpora, for tasks such as semantic tagging of nominal heads, semantic clustering of verbs, selectional preferences of (classes of) predicates.

These techniques will be applied to the selection of, navigation through and translation of multi-lingual information available through telematic systems and services. An extension to this application will add speech-driven access to the system.

Performance of the systems will be evaluated by standard criteria of information retrieval, against a baseline system which does not use phrasal or lexical information developed in the project.

6. A Generalization

The conclusion we would like to draw is based on the various experiences briefly outlined above, which should be taken as a framework in which we could also insert our future work strategy in the vast field of lexical/textual resources. This model strategy is represented in the figure below:

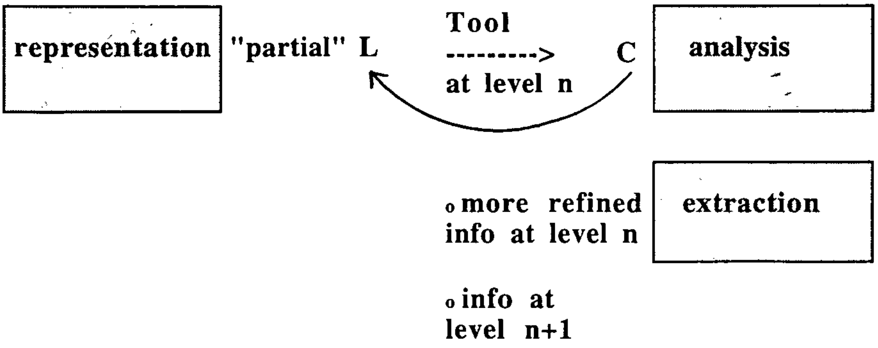

Figure 3. Future work strategy in the field of lexical/textual resourses

The lexicon (L) and the corpus (C) are related by a double association, thus implementing a cycle. In order to avoid this becoming a loop, we need to work with a bootstrapping methodology. We begin with a necessary "partial" (both in breadth and in depth) lexicon and we use it to perform some analysis of the corpus, using appropriate tools. The information present in the lexicon, and formally represented here, is usually projected on the corpus and reflected by the annotation/analysis performed on it. At this stage, one consequence is that we have an "enriched" corpus, from where – through usage of other tool types – we can acquire either more refined information at the same level of analysis or, more interestingly, information pertinent at another, superior, level of analysis. These information types can now be fed back into our lexicon which, still partial, is however richer and therefore able, in a successive implementation of the cycle, to achieve a richer analysis of the corpus.

Obviously what is said here, in a very simplistic way, needs a rather complex set of knowledge, tools, techniques, etc., in order to be performed. It is only recently that these resources are becoming robust enough to allow a number of cycles to be implemented with the hope of success. In particular, the capability of "integrating" what until now was used in isolation is essential: by using a good integration process the value of the individual pieces can be multiplied.

It is worth making a last observation: implementation of such a cycle needs a clear and strong compatibility both i) between the lexical representation and the corpus annotation, and ii) at the system/tools interface level (for input/output). From this consideration a clear need for continuing the standardisation efforts in language engineering emerges.

References

Calzolari, N. (1982): "Towards the organisation of lexical definitions on a database structure", in E. Hajicova (ed.) *COLING '82*, Charles University, Prague.

Calzolari, N. (1991): "Lexical databases and textual corpora: perspectives of integration for a Lexical Knowledge Base", in U. Zernik (ed.), *Lexical Acquisition: Using on-line resources to build a lexicon,* Erlbaum Ass., Hillstate, New York.

Calzolari, N. (forthcoming): "Observation and Generalisation: Corpus-based linguistic analysis of Italian speech-act verbs", in Yael (ed.), Oxford University Press.

Calzolari, N., Baker, M. and Kruyt, T.(eds., 1995): "Towards a Network of European Reference Corpora: Report of the NERC Consortium

Feasibility Study", in *Linguistica Computazionale XI - XII*, Giardini, Pisa.

Calzolari, N. and Briscoe, T. (1995): "ACQUILEX-I and -II: Acquisition of Lexical Knowledge from Machine-Readable Dictionaries and Text Corpora", in *Cahiers de Lexicologie, vol. LXVII, n.2.*

Calzolari, N., Federici, S., Montemagni, S. and Peters, C. (1994): "Extracting, representing and using syntactic-semantic information from Cobuild definitions", in J. Sinclair, M. Hoelter, C. Peters (eds.), *The Language of Definition: the formalisation of dictionary definitions for Natural Language Processing,* European Commission, Luxembourg.

Calzolari, N., Hagman, J., Marinai, E., Montemagni, S, Spanu, A. and Zampolli, A. (1993): "Encoding lexicographic definitions as Typed Feature Structures", in E. Beckmann, G. Heyer (Eds:), *Theorie und Praxis des Lexicons,* de Gruyter, Berlin.

Calzolari, N. and McNaught, J. (1996): "EAGLES Final Report: Editors' Introduction", *EAGLES document EAG-EB-EI,* Pisa.

Federici, S. and Pirrelli, V. (1994): "Linguistic Analogy as a Computable Process", in *Proceedings of NeMLaP,* Manchester, UK, pp. 8–14.

Fillmore, C.J. and Atkins, B.T. (1992): "Towards a frame-based lexicon: the semantics of RISK and its neighbours", in A. Leher, E. Kittay (eds.), *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organisation,* Erlbaum Ass., Hillstate, New York.

Heid, U. (1996): "A reading guide to the documentation produced by the EAGLES workgroup on computational lexicons", *EAGLES document EAG-LWG-Guide*, Stuttgart.

Heid, H. and McNaught, J. (eds., 1991): "Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications", *Eurotra-7,* Stuttgart.

King, M. (1996): "EAGLES Evaluation of Natural Language Processing Systems Report", *EAGLES document EAG-EWG-NLPS,* Copenhagen.

Monachini, M. and Calzolari, N. (1996): "Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages", *EAGLES document EAG-LWG-morphsyn,* Pisa.

Montemagni, S. and Pirrelli, V. (1995): "Do Lexical Rules Apply Across the Board? A corpus-based Investigation in the Machinery of Causative-Inchoative Alternation in Italian", In *Proceedings of the Acquilex Workshop on Lexical Rules*, Cambridge.

Resnik P.S., (1993): *Selection and Information: a Class-based Approach to Lexical Relationships*, PhD Dissertation, University of Pennsylvania.

Sparkle (Shallow PARsing and Knowledge extraction for Language Engineering) (1995): *Technical and Financial Annex for LE1-2111,* Luxembourg.

A. Zampolli (1995): “Introduction”, in N. Calzolari, M. Baker, and T. Kruyt (eds., 1995).