Standards and Best Practice for Multilingual Computational Lexicons & MILE (the Multilingual ISLE Lexical Entry)

Deliverable D2.2-D3.2
ISLE Computational Lexicon Working Group

Responsible Authors

Nicoletta Calzolari	Istituto di Linguistica Computazionale, CNR, Pisa, ITALY Consorzio Pisa Ricerche, Pisa, ITALY
Francesca Bertagna	Istituto di Linguistica Computazionale, CNR, Pisa, ITALY Università di Pisa, Pisa, ITALY
Alessandro Lenci	Università di Pisa, Pisa, ITALY
Monica Monachini	Istituto di Linguistica Computazionale, CNR, Pisa, ITALY

Authors

Sue Atkins[1], Nuria Bel[2], Pierrette Bouillon[3], Thatsanee Charoenporn[4], Dafydd Gibbon[5], Ralph Grishman[6], Chu-Ren Huang[7], Asanee Kawtrakul[4], Nancy Ide[8], Hae-Yun Lee[9], Paul J. K. Li[7], Jock McNaught[10], Jan Odijk[11], Martha Palmer[12], Valeria Quochi[13 & 14 & 15], Ruth Reeves[6], Dipti Misra Sharma[16], Virach Sornlertlamvanich[17], Takenobu Tokunaga[18], Gregor Thurmair[19], Marta Villegas[2], Antonio Zampolli[13 & 14 & 15], Elizabeth Zeiton[7]

1: Word Trade Centre, Lewes, UK. 2: gilcUB, Barcelona, SPAIN.
3: ISSCO, University of Geneva, Geneva, SWITZERLAND. 4: Kasetsart University, THAILAND.
5: Universität Bielefeld, Bielefeld, GERMANY. 6: New York University, New York, NY, USA.
7: Academia Sinica, Taipei, TAIWAN. 8: Vassar College, New York, NY, USA.
9: KORTERM, KOREA. 10: UMIST, Manchester, UK.
11: University of Utrecht, UiL-OTS, Utrecht, THE NEDERLANDS.
12: CIS Department, University of Pennsylvania, Philadelphia, PA, USA.
13: Consorzio Pisa Ricerche, Pisa, ITALY. 14: Istituto di Linguistica Computazionale, CNR, Pisa, ITALY.
15: Università di Pisa, Pisa, ITALY. 16: IIIT, INDIA. 17: NECTEC, THAILAND.
18: Tokyo Institute of Technology, JAPAN. 19: Comprendium, Munich, GERMANY

Table of Contents

1 Executive Summary

The ISLE Computational Lexicon Working Group (CLWG) is committed to the consensual definition of a standardized infrastructure to develop multilingual resources for HLT applications, with particular attention to the needs of Machine Translation (MT) and Crosslingual Information Retrieval (CLIR) systems. Compared with other standardization initiatives active in this field (e.g. OLIF-2), the original character of ISLE resides in its specifically focusing on the *gray area* of HLT where well-assessed language technology meets more advanced levels and forms of linguistic description. The ISLE CLWG aims at selecting mature areas and results in computational lexical semantics and in multilingual lexicons, which can also be regarded as stabilized achievements, thus to be used as the basis for future research.

For *multilingual computational lexicons*, ISLE objectives are: extending EAGLES work on lexical semantics, necessary to establish inter-language links; designing and proposing standards for multilingual lexicons; developing a prototype tool to implement lexicon guidelines and standards; creating exemplary EAGLES-conformant sample lexicons and tagging exemplary corpora for validation purposes; and developing standardized evaluation procedures for lexicons.

In particular, the ISLE-CLWG pursues this goal by designing MILE (*Multilingual ISLE Lexical Entry*), a general schema for the encoding of multilingual lexical information. This has to be intended as a meta-entry, acting as a common representational layer for multilingual lexical resources. Obviously ***MILE also includes previous EAGLES recommendations for other layers***.

Finally, one of the targets of standardization, and actually one of the main aims of the CLWG activities, is to create a common parlance among the various actors (both of the scientific and of the industrial R&D community) not only in the field of computational lexical semantics and multilingual lexicons, *but also in the areas e.g. of ontologies and the emerging semantic web*, so that synergies will be enhanced, commonalties strengthened, and resources and findings usefully shared. In other terms, the process of standard definition undertaken by CLWG, and by the ISLE enterprise in general, on one side 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 will actually be achieved. On the other side ***ISLE results pave the way to a needed cooperation between until now separate communities, such as HLT actors and groups specifically involved with 'content' (ontologies, semantic web, content providers, etc.)***, enabling future common efforts and resource sharing.

One of the first objectives of the CLWG was to discover and list the (maximal) set of (granular) *basic notions* needed to describe the multilingual level. Since a substantial part of the basic notions should be already included in previous EAGLES recommendations, and, with different distribution, in the existing and surveyed lexicons, and since the multilingual layer depends on monolingual layers, we had to revisit earlier linguistic analysis (previous EAGLES work, essentially monolinguistic) to see what we need to change/add or what we can reuse for the multilingual layer. *Sense distinctions are especially important for multilingual lexicons, since it is at this level that cross-language links need to be established.* The same is true of syntagmatic/collocational/ contextual information. To these areas we have paid particular attention in the recommendation phase, and we have examined how to extend the available EAGLES guidelines in these and other areas to propose a broad format for multilingual lexical entries which is of general utility to the community.

The principle guiding the elicitation and proposal of MILE basic notions in the recommendation phase has been, according to a previous EAGLES methodology, the so-called *'edited union'* (term put forward by Gerald Gazdar in earlier EAGLES work) of what exists in major lexicons/models/dictionaries, at least as a starting point, enriched with those types of information which are usually not handled, e.g. those of collocational/syntagmatic nature, and obviously those pertinent to the multilingual layer. This method of work has proven useful in the process of reaching consensual *de facto* standards in a bottom-up approach and is at the basis also of ISLE work. There is every interest in building on existing resources, rather than starting from scratch, thus efforts must continue in this direction.

In its general design, MILE is envisaged as a highly *modular* and *layered*. Modularity concerns the "horizontal" MILE organization, in which independent and yet linked modules target different dimensions of lexical entries. On the other hand, at the "vertical" level, a layered organization is necessary to allow for different degrees of granularity of lexical descriptions, so that both "shallow" and "deep" representations of lexical items can be captured. This feature is particularly crucial in order to stay open to the different styles and approaches to the lexicon adopted by existing multilingual systems.
This way, both at the monolingual and at the multilingual level (but with particular emphasis on the latter), ISLE intends to start up the incremental definition of a more Object-Oriented layer for lexical description, and to foster the vision of open and distributed lexicons, with elements possibly residing in different sites of the web. The defined lexical objects will be used by the lexicon (or applications) developers to build and target lexical data at a higher level of abstraction. Thus, they have to be seen as a step in the direction of simplifying and improving the usability of the MILE recommendations.

Not only will computational lexicons contribute to the content-based management of information on the Web, but the tools and resources that are being developed for the Semantic Web also provide the ground for the architecture and design of next-generation language resources. Moreover, computational lexicons should be conceived as *dynamic systems*, whose development needs to be complemented with the automatic acquisition of semantic information from texts. Gaining insights into the deep interrelation between representation and acquisition issues is likely to have significant repercussions on the way linguistic resources will be designed, developed and used for applications in the years to come. As the two aspects of knowledge representation and acquisition are profoundly interrelated, progress on both fronts can only be achieved, in our view of things, through a full appreciation of this deep interdependency.

2 Introduction

2.1 The EAGLES/ISLE Enterprise

ISLE[1] (*International Standards for Language Engineering*), a transatlantic standards oriented initiative under the Human Language Technology (HLT) programme, is a continuation of the long standing European EAGLES initiative (Calzolari, Mc Naught and Zampolli, 1996), carried out through a number of subsequent projects funded by the European Commission (EC) since 1993 (coordinated by A. Zampolli for the Consorzio Pisa Ricerche). EAGLES stands for *Expert Advisory Group for Language Engineering Standards* and was launched within EC Directorate General XIII' s Linguistic Research and Engineering (LRE) programme, continued under the Language Engineering (LE) programme, and under the Human Language Technology (HLT) programme as ISLE, since January 2000. ISLE is carried out by European and American groups within the EU-US International Research Co-operation, supported by EC and NSF. ISLE was built on joint preparatory EU-US work of the previous two years aimed at setting up a transatlantic standards oriented initiative for HLT.

The objective of the project is to support HLT R&D international and national projects, and industry by developing, disseminating and promoting widely agreed and urgently demanded HLT standards and guidelines for infrastructural language resources (see Zampolli, 1998, and Calzolari, 1998), tools that exploit them, and LE products. The aim of EAGLES/ISLE is thus to accelerate the provision of standards, common guidelines, best practice recommendations for:

- very large-scale language resources (such as text corpora, computational lexicons, speech corpora (Gibbon *et al.*, 1997), 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 (EAGLES, 1996).

Leading industrial and academic players in the HLT field (more than 150) have actively participated in the definition of this initiative and have lent invaluable support to its execution over the years. Moreover, the initiative is a direct result of a series of recommendations made to the EC over several years. There is a recognition that standardization work is not only important, but is a necessary component of any strategic programme to create a coherent market, which demands sustained effort and investment.

It is important to note that the work of EAGLES (see EAGLES guidelines, http://www.ilc.pi.cnr.it/EAGLES96/home.html) must be seen in a long-term perspective. Moreover, successful standards are those which respond to commonly perceived needs or aid in overcoming common problems. In terms of offering workable, compromise solutions, they must be based on some solid platform of accepted facts and acceptable practices. EAGLES was set up 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 language technology, through bringing together representatives of major collaborative European R&D projects, and of HLT industry, in relevant areas. This work has been conducted with a view to providing the foundation for any future recommendations for International Standards that may be formulated under the aegis of ISO.

[1] ISLE Web Site URL: lingue.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

The ISLE project (coordinated by A. Zampolli for EU and M. Palmer for US) (see http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm) targets the three areas of

- *multilingual computational lexicon* (EU chair: N. Calzolari; US chairs: M. Palmer and R. Grishman),
- *natural interaction and multimodality* (*NIMM*) (EU chair: N. O. Bernsen; US chair: M. Liberman),
- *evaluation of HLT systems* (EU chair: M. King; US chair: E. Hovy).

These areas were chosen not only for their relevance to the HLT field but also for their long-term significance. Three Working Groups, and their sub-groups, have carried out the work, according to the already proven EAGLES methodology, with experts from both the EU and US, working and interacting within a strongly co-ordinated framework. Responsible partners recruit members from the HLT community (from both academia and industry) to participate in working groups. International workshops are used as a means of achieving consensus and advancing work. Results are widely disseminated, after due validation in collaboration with EU and US HLT R&D projects, National projects, and industry.

In the following we concentrate on the *Computational Lexicon Working Group (CLWG)*.
The deliverable is the result of the activity of two work packages devoted, on one hand, to the extension of the previous EAGLES work on Lexical Semantics and on the other hand to the actual recommendations and guidelines for the representation of information in bi-multilingual lexica.
The deliverable can be conceived as a two-part document: the first *macrosection* (covering the chapters 1 to 5) concerns the general description of the project and the "background" activity devoted to linguistic/lexicographic analysis aimed to the identification of types of information that play a crucial role in a bi-multilingual environment.
In particular, **Chapter 3** describes EAGLES/ISLE CLWG specific methodology and its goal of establishing a general and consensual standardized environment for the development and integration of multilingual resources.
Chapter 4 briefly presents the first phase of activities of the CLWG, dedicated to the elaboration of a survey of existing multilingual resources both in the European, American and (although still in a more limited extension) Asian research and industrial scenarios. Such a review is also the basis for the process of standard selection and definition, which was the focus of the second phase of the CLWG, aiming at individuating hot areas in the domain of multilingual lexical resources, which call – and *de facto* can access to – a process of standardization. A specific deliverable (Calzolari et al., 2001a) is dedicated to the survey phase. In Chapter 4 a brief description of the two major background lexical resources is also provided: PAROLE/SIMPLE semantic lexicons built on the basis of previous EAGLES recommendations, and WordNet lexicons.
Chapter 5 is dedicated to the results of two different (yet interlinked) lexicographic studies leading to the identification of the information that plays a role, at any level of linguistic description, in the multilingual transfer: section 5.1 deals with a concrete entry creation activity (about 100 multilingual entries for French, Italian and English), section 5.2 describes the use of the so-called *sense indicators* in machine readeble dictionaries as candidates for transfer conditions. To the work of entry creation a specific deliverable has been dedicated (the D5.1).
The second macrosection goes from **Chapter 6** onwards, where we describe the ISLE proposals of guidelines for the "Multilingual ISLE Lexical Entry" (MILE), the general architecture and features of MILE, as well as discuss the methodology adopted for its definition. The general architecture of MILE is presented in section 6.1. Then, we focus on the two essential ingredients for the MILE specification: on the one hand the selection of the types of lexical information most relevant to establish multilingual correspondences, the MILE basic notions (section 6.2), and on the other hand the specification of the MILE linguistic data model, which will provide the formal backbone of the

MILE as a general representation language to develop multilingual resources and to link computational lexicons (6.3). The formalization of the MILE in RDF is given in section 6.4.
The description of the MILE prototype tool to implement lexicon guidelines and standards can be found in **Chapter 7.** The tool is a lexicographical station development platform in order to automatically map the DTD into a relational dB, build up a user-friendly interface able to cover the most common requirements of a lexicographic station and to exemplify, test and validate the goodness of the MILE model in a real scenario, that is, reusing already existing monolingual resources such as PAROLE and SIMPLE lexicons. Chapter 8 briefly at future strategies for Language Resources building on ISLE results.

Other important sections of the deliverable are the appendices, dedicated to encoding examples, case studies, exemplificative data, etc.
Appendix A, **B** and **C** are dedicated to RDF formalization of the MILE: **Appendix A** presents an RDF Schema for MILE, **Appendix B** exemplifies the syntactic layer for some lexical entries, and **Appendix C** defines lexical objects of the *Lexical Data Category Registry* (LDCR).
In **Appendices D** and **E** two important reference ontologies are presented, respectively the one from SIMPLE and the EuroWordNet Top Ontology.
One of the important issues ISLE has to deal with is the problem of the representation of noun compounds and support verbs in MILE: a study on the subject, conducted in collaboration with the XMELLT[2] project, is presented in **Appendix F**.
A crucial aspect in establishing a real and broad consensus is played by communication and sharing of information among many groups active in the field. For this reason we involved also Asian colleagues in the ISLE initiative, and we have explored ways of establishing formal links with them. **Appendix G** provides an exemplification of this collaboration, showing how the EAGLES/ISLE basic notions have been "tested" in the representation of some Asian languages.
Another important issue is the broadening of the scope of ISLE towards a more general and comprehensive description and representation of the linguistic "universe": in **Appendix H** a survey of the issues that are distinctive of the Spoken in respect of Written Language is presented. This can considered a first, important step towards the integration of the ISLE Basic Notions with information important for "wider" multimodal aspects.

[2] "Cross-lingual Multiword Expression Lexicons for Language Technology", N. Ide, Vassar, PI, NSF Award No. 9982069, May 1, 200- Dec. 31, 2001.

3 The ISLE Computational Lexicon Working Group: Objectives and Methodology

3.1 Objectives of Work

For *multilingual computational lexicons*, ISLE objectives are: i) extending EAGLES work on lexical semantics, necessary to establish inter-language links; ii) designing and proposing standards for multilingual lexicons; iii) developing a prototype tool to implement lexicon guidelines and standards; iv) creating exemplary EAGLES-conformant sample lexicons and tagging exemplary corpora; and v) developing evaluation procedures for lexicons.

The ISLE Computational Lexicon Working Group (CLWG) is committed to the *consensual definition of a standardized infrastructure to develop multilingual resources for HLT applications*, with particular attention to the needs of Machine Translation (MT) and Crosslingual Information Retrieval (CLIR) systems. Compared with other standardization initiatives active in this field (e.g. OLIF-2; cf. Lieske *et al.*, 2001), the original character of ISLE resides in its specifically focusing on the *gray area* of HLT where well-assessed language technology meets more advanced levels and forms of linguistic description. With no intent of imposing any constraints on investigation and experimentation, the ISLE CLWG rather aims at selecting mature areas and results in computational lexical semantics and in multilingual lexicons, which can also be regarded as stabilized achievements, thus to be used as the basis for future research.

In particular, various aspects of lexical semantics, although still part of ongoing research, are nevertheless regarded by industrials and developers as the "next-step" in new generation multilingual applications. Lexical semantics has always represented a sort of *wild frontier* in the investigation of natural language, let alone when this is also aimed at implementing large scale systems based on HLT components. In fact, the number of open issues in lexical semantics both on the representational, architectural and content level might induce an actually unjustified negative attitude towards the possibility of designing standards in this difficult territory. Rather to the contrary, standardization must be conceived as enucleating and singling out the areas in the open field of lexical semantics, that already present themselves with a clear and high degree of stability, although this is often hidden behind a number of formal differences or representational variants, that prevent the possibility of exploiting and enhancing the aspects of commonality and the already consolidated achievements. Standard definition in this area thus means to lay a first bridge between research in multilingual resource development and its exploitation in advanced technological systems.

The ISLE CLWG pursues this goal by designing the **MILE (*Multilingual ISLE Lexical Entry*)**, a general schema for the for encoding of lexical information multilingual linking. This has to be intended as a meta-entry, acting as a common representational layer for multilingual lexical resources.

Consistently, the ISLE standardization process pursues a twofold objective:

1. defining standards both at the content and at the representational level for those aspects of computational lexicons which are already widely used by applications;
2. proposing recommendations for the areas of computational lexical semantics which are still in the "front line" of ongoing research, but also appear to be ready for their applicative

exploitation, and are most required by HLT systems to achieve new technological leap forwards.

ISLE Standardization effort has to provide a way to deal with both symbolic and sub-symbolic (statistical) model of tools and applications.

This multidimensional perspective is one of the peculiar features of the ISLE activities, and contributes to its added value with respect to other current standardization initiatives. This way, ISLE intends on the one hand to answer to the need of fostering the reuse and interchange of existing lexical resources, and on the other hand to enhance the technological transfer from advanced research to applications. It also prepares the ground for a "new generation" of "knowledge resources".

3.2 Standards design and the interaction with R&D

EAGLES work towards *de facto* standards has already allowed the field of Language Resources (LR) to establish broad consensus on key issues for some well-established areas — and has allowed similar consensus to be achieved for other important areas through the ISLE project — providing thus a key opportunity for further consolidation and a basis for technological advance. EAGLES previous results in many areas have in fact already become *de facto* widely adopted standards, and EAGLES itself is a well-known trademark and a point of reference for HLT projects and products.

Existing EAGLES results in the Lexicon and Corpus areas are currently adopted by an impressive number of European - and recently also National - projects, thus becoming "the *de-facto* standard" for LR in Europe. They are now evaluated to become a basis for Asian Language Resources (LR) specifications. 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, Zampolli, 1997, Ruimy *et al.*, 1998, Lenci *et al.*, 1999, Bel *et al.*, 2000) rely on EAGLES results (Sanfilippo, A. *et al.*, 1996 and 1999), and are now being enlarged at the national level through many National Projects;
- the ELRA Validation Manuals for Lexicons (Underwood and Navarretta, 1997) and Corpora (Burnard *et al.*, 1997) are based on EAGLES guidelines;
- morpho-syntactic encoding of lexicons and tagging of corpora in a very large number of EU, international and national projects – and for more than 20 languages — is conformant to EAGLES recommendations (Monachini & Calzolari, 1996, 1999, Leech and Wilson, 1996).
- experiment dedicated to the application of the EAGLES/ISLE *basic notions* in the representation of some Asian languages (cf. Appendix G).

The fact that the core PAROLE/SIMPLE resources are now enlarged to real-size lexicons within National Projects in at least 8 EU countries allows to have a really large infrastructural platform of harmonised lexicons in Europe, sharing the same model.

For a standardization initiative it is however important also to accept and incorporate *de facto* standards which have imposed themselves in the LR community. This is the case of e.g. the EuroWordNet lexicons, now available for many EU languages, or the OLIF standards, used by a number of industrial multilingual systems. ISLE takes this into account.

Standards must emerge from state-of-the-art developments. With this respect, 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 involves many bodies active in EU-US NLP and speech projects, close collaboration with these projects is 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. Procedures have been established allowing EAGLES to access relevant material developed by EAGLES participants working in other projects. As an example, the NSF project XMELLT on multi-words for multilingual lexicons has provided valuable input to ISLE.

The consolidation of a standards proposal must be viewed, by necessity, as a slow process comprising, after the phase of putting forward proposals, a cyclical phase involving EAGLES 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 European 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 arrived at;
- dissemination and promotion of consensual recommendations;
- promotion of the standard to become International ISO (International Organization for Standardization, available at: http://www.iso.ch/) Standard. With respect to ISO, coordination with ISO and promotion of the possibility of adopting EAGLES/ISLE standard as a basis for ISO is ensured by the participation to ISO Committee for LR.[3]

What can be defined as *new advance* in this process is the highlighting of the areas for consensus (or of the areas in which consensus could be reached) and the gradual consciousness of the stability that evolves within the communities involved. A first benefit is the possibility, for those working in the field, of focusing their attention on as yet unsolved problems without losing time in rediscovering and re-implementing what many others have already worked on. This is the only way our discipline can really move forward.

3.3 Scope of the recommendations and type of users

The basic idea behind EAGLES work has always been for the group to act as a catalyst in order to pool concrete results coming from current major International/National/industrial projects. Relevant common practices or upcoming standards are used where appropriate as input to EAGLES/ISLE work. Numerous theories, approaches, and systems are being taken into account, where appropriate, as any recommendation for harmonization must take into account the needs and nature of the different major contemporary approaches. EAGLES is also drawing strong inspiration from the results of major projects whose results have contributed to advancing our understanding of harmonization issues. A strong characteristic of EAGLES has always been its openness to new developments in the field and its capability of integrating innovative and emerging directions.

The major efforts in EAGLES concentrate on the following types of activities, which, as seen in the following, show how, on very general lines, the work is organised in the working groups.

- Detecting those areas ripe for short-term standardization vs. areas still in need of basic research and development;
- 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 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;
- Setting up guidelines for representation of core sets of basic features, for representation of resources, etc.;
- Feasibility studies for less mature areas;

[3] Antonio Zampolli is member of the ISO Advisory Board and Nicoletta Calzolari is member of the Committee.

- Suggesting actions to be taken for a stepwise procedure leading to the creation of multilingual reusable resources, elaboration of evaluation methodologies and tools, etc.
- Paving the way to innovative types of resources.

The general vision adheres to the idea of enhancing the sharing and reusability of multilingual lexical resources, by promoting the definition of a common parlance for different communities (both scientific and industrial R&D communities): not only in the field of multilingual HLT and computational lexicon developers, *but also in the areas e.g. of ontologies and the emerging semantic web*, so that synergies will be enhanced, commonalties strengthened, and resources and findings usefully shared. In other terms, the process of standard definition undertaken by CLWG, and by the ISLE enterprise in general, on one side 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 will actually be achieved. On the other side *ISLE results pave the way to a needed cooperation between until now separate communities, such as HLT and other actors and groups specifically involved with 'content' and knowledge (ontologies, semantic web, content providers, etc.)*, enabling future common efforts and resource sharing.
Another critical strategy for future use of the ISLE recommendations is the recognition of the importance for the recommendations to be such that can be used both for information of rich lexicons and e.g. for simple domain specific lexicons with not much linguistic information. This also enables ISLE results to be used by a very large spectrum of LR and Knowledge Resources builders, within the same general model and architecture.

The design of a common and standardized framework for lexicon and knowledge resources construction can lead to the optimization of the whole process of production of resources: their creation, maintenance and (also automatic) extension, but also their reusability for different applications and tasks. It is critical to achieve the interoperability needed for effective integration, a precondition for a qualitative improvement in multilingual content processing technologies and for the Semantic Web vision.

3.4 ISLE Methodology and Organization of Work

In the process of specifying the various components of MILE, the ISLE-CLWG has adopted a three-step methodology:

i) survey of existing monolingual/multilingual lexicons and best practices in computational lexicography as applied in HLT (cf. section 4 and Calzolari et al., 2001a);
ii) identifying the lexical dimensions and the various types of information which are relevant to establish multilingual correspondences. These have been termed *basic notions for multilingual lexical encoding*. The selection of the basic notions has involved a twofold lexicographic in-depth investigation. The latter has consisted of an experiment of intensive lexical entry writing, paired with an analysis of common practice in multilingual lexicography;
iii) defining a suitable formal data model to encode the basic notions as well as the operations required at the multilingual level.

4 The ISLE Survey Phase and Background References

4.1 Survey of major computational lexicons

Following the well established EAGLES methodology, the first priority of the CLWG in the first phase of the ISLE project was the drafting of a wide-range survey of bilingual/multilingual (or semantic monolingual) lexicons, so as to reach a fair level of coverage of existing lexical resources.

This phase was a preliminary and yet crucial step towards the main goal of the CLWG, i.e. the definition of the "*Multilingual ISLE Lexical Entry*" (*MILE*). This was the main focus of the second phase of the project, the so called "recommendation phase", whose main objective was proposing consensual Recommendations/Guidelines.

With respect to this target, one of the first objectives of the CLWG is to discover and list the (maximal) set of (granular) *basic notions* needed to describe the multilingual level. Since a substantial part of the basic notions should be already included in previous EAGLES recommendations, and, with different distribution, in the existing and surveyed lexicons, and since the multilingual layer depends on monolingual layers, we had to revisit earlier linguistic analysis (previous EAGLES work, essentially monolinguistic) to see what we need to change/add or what we can reuse for the multilingual layer.

4.1.1 The Survey Phase

The *Survey* of existing lexicons (see Calzolari, Grishman, Palmer, eds. 2001) has been accompanied by the analysis of the requirements of a few multilingual applications, and by the parallel analysis of typical cross-lingually complex phenomena. Both these aspects have provided the general scenarios in terms of which the survey has been organized and carried out, as well as they have been the reference landmarks for the proposal phase of standard design.

The function of an entry in a multilingual lexicon is to supply enough information to allow the system to identify a distinct sense of a word or phrase in the Source Language (SL), in many different contexts, and reliably associate each context with the most appropriate translation in the Target Language (TL). The main issue is how to state in the most proper way the restrictions that both in SL or/and in TL make the translational relation true. This was referred as "transfer conditions" in early MT literature and has now spread for cross-lingual and multilingual information handling. In addition, the passage from SL to TL makes it necessary to express how information is passed from one language to another taking into account as difficult and pervasive phenomena as head-argument differences, relevance of collocational patters and multi-word expressions, etc.

The first step is to determine, of all the information that can be associated with SL lexical entries, what is the most relevant to a particular task, e.g. which notions are the more relevant to be encoded, at which descriptive level, to which elements of the entry conditions and actions for translation need to be associated, etc. The following is a (non-exhaustive) list of key applications which rely on the use of lexical resources:

– Machine Translation (MT)

– Cross-Language Information Retrieval (CLIR)
– Cross-Language Information Extraction
– Multilingual Language Generation
– Multilingual Authoring
– Speech-to-Speech Translation
– Multilingual Summarization

We decided to *focus the work of survey and subsequent recommendations around two major broad categories of application: MT and CLIR.* They have partially different/complementary needs, and can be considered to represent the requirements of other application types. The multilingual applications, considered as a starting point for both phases, provide a strong applied focus in tackling multilingual lexical encoding. It is necessary in fact to ensure that any guidelines meet the requirements of industrial applications and that they can be implemented.

In the preparation of the Survey, both to facilitate the identification of basic notions and the comparison of surveyed resources, and to focus on aspects of relevance to multilingual tasks, we have decided:

1. to prepare a grid for lexicon description to be used as a checklist to classify the content and structure of the surveyed resources on the basis of a number of agreed parameters of description;
2. to identify a small number of major categories of cross-lingual lexical phenomena that could be used to focus the survey, and to provide the necessary bootstrap to the proposal phase. Actually, they represent typical *hard cases,* which are helpful to highlight the various strategies that different lexicons and systems typically resort to when operating in multilingual environments. It is one of the expected by-products of the global CLWG activity to extend and refine this preliminary list, so as to provide researchers and developers with an updated map of the problematic cases in the realm of lexical information formalization, storage, and access, together with proposals on how to tackle them.

In order to better analyze lexicons, we organized the Survey in three different types of resources:

- *Machine Readable Dictionaries (MRDs),* where the rich monolingual and bilingual information is typical of the lexicographic tradition;
- *Computational Lexicons,* large lexical resources for general use where detailed morphosyntactic, syntactic and semantic information is explicit and variously represented;
- *Lexical resources for Machine Translation systems.*

Each lexicon presentation includes:

a. a description of the surveyed resource (also on the basis of the common grid, see Table 1);
b. possibly, for one or two examples from the cross-lingual lexical phenomena, an explanation of how these examples are handled by this lexicon.

The following template (drawn up starting from a preliminary list that essentially concerned the information present in traditional dictionaries, then integrated with more detailed morpho-syntactic, syntactic and semantic information, which might be available in existing computational lexicons and machine-readable dictionaries) has been used as a general grid to evaluate the content and structure of each lexical resource by i) verifying if the information is available and extractable, and ii) focusing on how the various types of information can be relevant to solve problems usually tackled when processing language in a bilingual or multilingual environment. The grid used in the survey phase is obviously not intended to be complete, since it is expected that new items might be introduced as a result of the recommendation phase.

	Entry component	
1	headword	
2	Phonetic transcription	
3	variant form	
4	inflected form	
5	Cross-reference	
6	Morphosyntactic information	
	a	Part-of-speech marker
	b	Inflectional class
	c	Derivation
	d	Gender
	e	Number
	f	Mass vs. Count
	g	Gradation
7	Subdivision counter	
8	Entry subdivision	
9	Sense indicator	
10	linguistic label	
11	Syntactic information	
	a	Subcategorization frame
	b	Obligatority of complements
	c	Auxiliary
	d	Light or support construction
	e	Periphrastic constructions
	f	Phrasal verbs
	g	Collocator
	h	Alternations
12	Semantic information	
	a	Semantic type
	b	Argument structure
	c	Semantic relations
	d	Regular polysemy
	e	Domain
	f	Decomposition
13	Translation	
14	Gloss	
15	near-equivalent	
16	example phrase (straightforward)	
17	example phrase (problematic)	
18	multiword unit	
19	Subheadword *also* secondary headword	
20	usage note	
21	Frequency	

Table 1: Lexical Information in Bilingual Resources

4.2 Background resources

4.2.1 The PAROLE/SIMPLE lexicons and the GENELEX model

Given the fact that the PAROLE/SIMPLE Lexicons (based on the GENELEX 1994 model) have been used and critically evaluated as a basis for the definition of the MILE, we briefly provide here some information of these resources. They cover 12 languages: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish.

All the PAROLE/SIMPLE lexical information is encoded in SGML (for Italian now also in XML), and the whole PAROLE/SIMPLE model is fully represented according to a common DTD for all the 12 languages, based on the GENELEX DTD (GENELEX Consortium, 1994).

The PAROLE lexical resources (Ruimy *et al.* 1998) encode the following morphological and syntactic information, divided into optional and mandatory classes for entries:

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

The design of the SIMPLE lexicons (Bel *et al.*, 2000) complies with the EAGLES Lexicon/Semantics Working Group guidelines (Sanfilippo *et al.*, 1999), and the set of recommended semantic notions. The SIMPLE lexicons (see http://www.ub.es/gilcub/SIMPLE/simple.html for the specifications and sample lexical entries for the various languages) are built as a new layer connected to the PAROLE syntactic layer, and encode structured "semantic types" and semantic (subcategorization) frames. Each lexicon is based on the same common model, designed to facilitate future cross-language linking: they share the same *core ontology* and the same set of *semantic templates*.

The SIMPLE model provides the formal specification for the representation and encoding of the following information:

i. *semantic type*, corresponding to the template that each Semantic Unit (*SemU*) instantiates;
ii. *domain* information;
iii. lexicographic gloss;
iv. *argument structure* for predicative SemUs;
v. selectional restrictions on the arguments;

vi. *event type*, to characterize the aspectual properties of verbal predicates;

vii. *links* of the arguments *to the syntactic subcategorization frames*, as represented in the PAROLE lexicons;

viii. *‘qualia’ structure*, following the Generative Lexicon (Pustejovsky, 1995), represented by a very large set of semantic relations and features;

ix. information about *regular polysemous alternation* in which a word-sense may enter;

x. information concerning cross-part of speech relations (e.g. intelligent - intelligence; writer - to write).

xi. *semantic relations*, such as hyponymy, synonymy, etc.

The “conceptual core” of the lexicons consists of the basic structured set of “semantic types” (the *SIMPLE ontology*) and the basic set of notions to be encoded for each sense. These notions have been captured in a common “library” of language independent *templates*, which act as “blueprints” for any given type - reflecting well-formedness conditions and providing constraints for lexical items belonging to that type.

1. TELIC [Top]
   …
2. AGENTIVE [Top]
   2.1. Cause [Agentive]
   …
3. CONSTITUTIVE [Top]
   3.1. Part [Constitutive]
      3.1.1. Body_part [Part]
   3.2. Group [Constitutive]
      3.2.1. Human_group [Group]
   3.3. Amount [Constitutive]
   …
4. ENTITY [Top]
   4.1. Concrete_entity [Entity]
      4.1.1. Location [Concrete_entity]
   …

Figure 1: A portion of the SIMPLE Ontology**.**

There are three main types of formal entities:

- *Semantic Units* – word-senses are encoded as *Semantic Units* or *SemU*. Each SemU is assigned a *semantic type* from the Ontology, plus other sorts of information specified in the associated *template,* which contribute to the characterization of the word-sense.

- *Semantic Type* - SemUs are assigned semantic types. Each type involves structured information represented as *template*. The semantic types themselves are organized into the *Ontology* (see Figure 1), which allows for the *orthogonal organization* of types (Pustejovsky, 1995). For a complete list of the SIMPLE Semantic Types, cf. Appendix D.

- *Template* - a schematic structure which the lexicographer uses to encode information about a given lexical item. The template expresses the semantic type, plus other sorts of information characterizing multiple dimensions of a word-sense. Templates are intended both to provide the semantics of the types (which are thus not simply labels) and to guide, harmonize, and facilitate the lexicographic work, as well as to enhance the consistency among the lexicons. A set of top common templates (about 150) for all the languages have been defined during the specification phase, while the individual lexicons can add more language-specific templates as needed.

Templates provide the information that is type-defining for a given semantic type. Lexicographers can also further specify the semantic information in a SemU, by either adding other relations or features in the Qualia Structure, or by adding other types of information (e.g. domain information, collocations, etc.).

Take, for instance, the template associated with the type `Instrument` (Table 2), followed by the SemU for a sense of *lancet*, instantiating this template (Table 3):

Usem:	1
BC number:	
Template_Type:	[Instrument]
Unification_path:	[Concrete_entity | Artifact$_{Agentive}$ | Telic]
Domain:	General
Semantic Class:	<Nil>
Gloss:	//free//
Event type:	<Nil>
Pred_Rep.:	<Nil>
Selectional Restr.:	<Nil>
Derivation:	<Nil>
Formal:	*isa* (1,<instrument>)
Agentive:	*created_by*(1,<Usem>:[Creation])
Constitutive:	*made_of*(1,<Usem>) //optional// *has_as_part*(1,<Usem>) //optional//
Telic:	*used_for*(1,<Usem>: [Event])
Synonymy:	<Nil>
Collocates:	*Collocates*(<Usem1>,..,<Usemn>)
Complex:	<Nil> //for regular polysemy//

Table 2: Instrument Template

Usem:	<lancet-1>
BC number:	
Template_Type:	[Instrument]
Unification_path:	[Concrete_entity| Artifact$_{Agentive}$ | Telic]
Domain:	Medicine
Semantic Class:	Instrument
Gloss:	a surgical knife with a pointed double-edged blade; used for punctures and small incisions
Event type:	<Nil>
Pred_Rep.:	<Nil>
Selectional Restr.:	<Nil>
Derivation:	<Nil>
Formal:	*isa* (<lancet-1>,<knife>:[Instrument])
Agentive:	*created_by*(<lancet-1>,<make>: [Creation])
Constitutive:	*made_of*(<lancet-1>,<metal>: [Substance]) *has_as_part* (<lancet-1>, <edge>: [Part])
Telic:	*used_for*(<lancet-1>,<cut>: [Constitutive_change]) *used_by* (<lance-1t>, <doctor>)
Synonymy:	<Nil>
Collocates:	*Collocates* (<SemU1>,..,<SemUn>)
Complex:	<Nil>

Table 3: SemU for *lancet*

4.2.2 The WordNet-type lexicons

During the years, WordNet (http://www.cogsci.princeton.edu/~wn/), the *electronic lexical database* that has been being developed at Princeton since the early 80' by G. Miller and his group, has become an outstanding reality for the lexicon community. Its architecture (together with its multilingual versions EuroWordNet) has already been described in the Survey of Available Lexicons (Calzolari et al., 2001a) so we will recall here only some basic principles.

WordNet, whose design derived by psycholinguistic and computational theories of human lexical memory, is a semantic network representing words and concepts as an interrelated system consistent with the evidence for the way speakers organize their mental lexicon. The *synset* is the set of synonyms that plays the central role of *lexical concept* in WordNet meshes, working as anchor for every semantic relations. The most important relation in WordNet is hyponymy/hyperonymy, which is the vertical backbone of the entire net by means of which each synset has to be anchored to a superordinate synset (up to the *unique beginners*, the synsets that are on the top of the hierarchy). Horizontally, WordNet has many other semantic relations. The following is a list of the relations available in the 1.5 release of the database:

- Antonymy
- Hyperonymy/ Hyponymy
- Meronymy/ Holonymy (member, substance, part)
- Entailment
- Cause
- Attribute
- Similarity

The current WordNet version, the 1.7, contains 111,223 synsets, divided in nouns, verbs, adjectives and adverbs; the relations link only synsets of the same part of speech except for Attribute, which links adjectives and nouns. WordNet is the model of many other similar lexical resources: in particular, its modality of representation of the lexical meaning converged in the design of EuroWordNet (http://www.illc.uva.nl/EuroWordNet/) , a system of eight monolingual semantic networks (for Dutch, Italian, Spanish, English, Czech, Estonian, French, German) linked by means of an Interlingua Index. Also within EuroWordNet, semantic information is encoded in each of the languages dealt with in form of lexical semantic relations between *synsets*. The set of semantic relations of WordNet was extended introducing new relations for their supposed relevance and usefulness in linguistic applications (e. g. cross part of speech relations)[4], but the most important factors of innovation of EWN in respect of the original WordNet are:

i) **Multilinguality**, reached via an Interlingual Index (ILI), i.e. an unstructured version of the Princeton WordNet1.5 containing all the synsets belonging to this version but not the relations among them. All the synsets of the monolingual modules of EuroWordNet are linked to this ‘interlingua” by means of a set of **relations of equivalence** to make the resource usable in multilingual applications. A subset of the ILI was circumscribed, in order to group together all the synsets considered basic concepts (Base Concept, BC) in each language. This subset, which is common to all the EWN languages, works as a means to link the language specific basic concepts to the language independent ontological structure.

ii) a **Top Ontology** (TO), a hierarchy of language independent concepts reflecting fundamental semantic distinctions and linked to all the monolingual modules via the set of Base Concepts.

[4] For the list of the 70 semantic relations available in EuroWordNet cf. Vossen, 1999.

The following picture (Figure 2) shows an example of the monolingual net surrounding the Italian synset {cane 1} (dog) and its relations of equivalence with the ILI. Dog is also linked, by means of the corresponding base concept, to the Top Concepts of the Ontology.

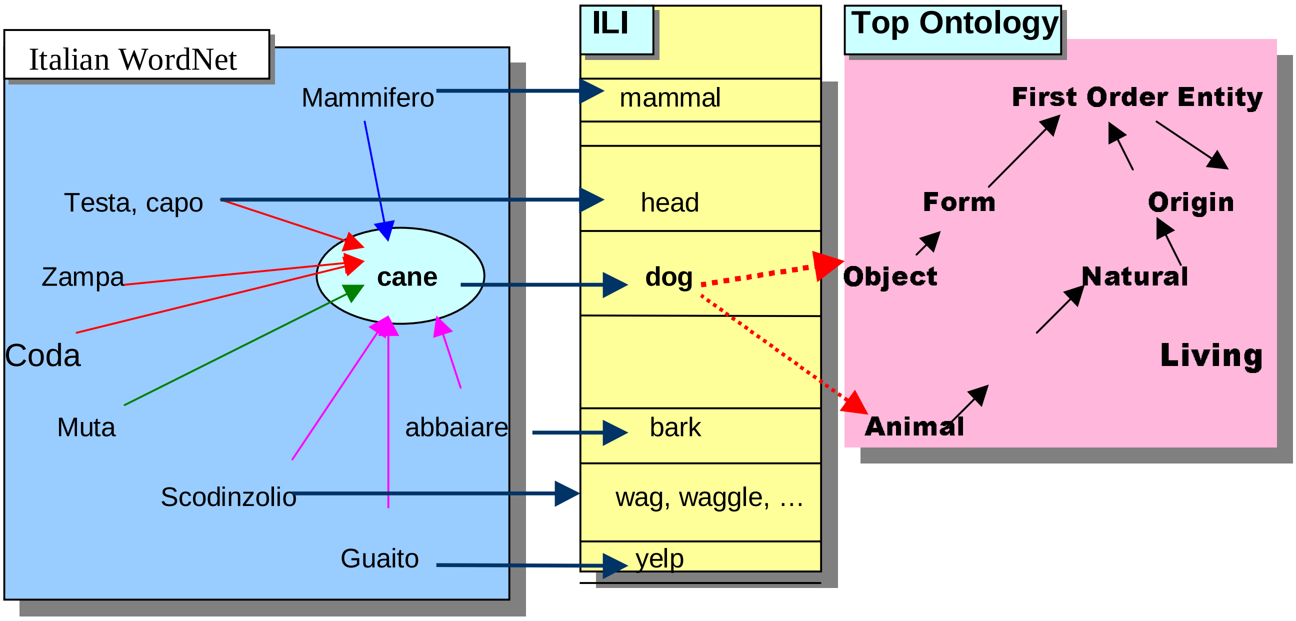


Figure 2: Portion of the ItalWordNet Lexicon for the synset {cane 1}

The resources developed within the EuroWordNet framework are not the only ones dedicated to languages other than English. Wordnets for dozens of languages have been built or are under development (for an updated list of them see: http://www.globalwordnet.org/gwa/wordnet_table.htm) and they are applied in a wide variety of applications.

Thus it is important to take WordNet and its basic structure into consideration, ensuring that all the already encoded resources could be easily mapped into the final ISLE recommendations.

5 Lexicographic Work Towards the ISLE *Basic Notions*

The purpose of this phase of the ISLE work has been to identify the lexical dimensions and the various types of information which are relevant to establish multilingual correspondences. These have been termed *basic notions for multilingual lexical encoding*.

To tackle this point, the survey of the available computational lexicons and system needs, carried out in the preliminary phases of the project (cf. Calzolari *et al.*, 2001a), has been complemented with a more lexicographic-based effort, to identify the types of information used in bilingual dictionaries to establish translation equivalents. To this purpose, the CLWG has organized two "task forces" with the responsibility respectively of creating a sample of lexical entries and investigating the use of *sense indicators* in traditional bilingual dictionaries. The aim of these activities has been twofold: i) highlighting the various types of information useful to determine the transfer conditions; ii) exploring and evaluating the full expressive potentialities provided by the reference computational model (i.e. the PAROLE-SIMPLE architecture).

5.1 Sample lexical entries

The task of "creating sample" of MILE-conformant entries has twofold motivations:

1. on one hand, at the beginning of the project, a case-study phase devoted to a lexicographic analysis was necessary, whose results have been of input for the recommendations.
2. on the other hand, when the MILE design was established, the entry creation has been important to "stretch" the potentialities of the model and to verify whether it was sufficiently flexible so as to permit the solving of lexicographic or translation problems.

In section 5.1.1 a description of the "case study" phase can be found while in section 5.1.2 we introduce the "testing" of the model in a real entry-encoding scenario.

5.1.1 The "case-study" phase

A first, small group of words (lemma) (the nouns *blow* and *day*, the verb *to hit* and *to play* and the adjective *round*) have been selected on the basis of their degree of polysemy and complexity of translation (and also because they were part of the set of entries used in the SENSEVAL-2 experiment), to build a general "test suite" of possible multilingual transfer scenarios. The aim of the case study was the selection of the information useful to link an L1 entry to its correspondent L2 lexical entry/ies, followed by the study of the possible ways of actually represent this information for multilingual purposes. This experiment has started with English and Italian, adopting the following procedure:

1. for each of the selected entries, we extracted the occurrences from various monolingual reference corpora (e.g. PAROLE- Corpus for Italian, BNC for English (Burnard 2001));
2. the extraction results have been organized in senses, with the help of existing monolingual dictionaries and computational lexicons (e.g. SIMPLE, WordNet, EuroWordNet/ItalWordNet, ComLex);

3. the relevant syntactic descriptions and the identified senses have been encoded according to the PAROLE-SIMPLE specifications (Lenci et al. 1999). The result was a core of monolingual lexical entries described at the morphological, syntactic and semantic levels;
4. the various identified senses have been translated using bilingual dictionaries, and the translations have been revised by native language speakers;
5. on the basis of (4), the monolingual entries have been linked into bilingual entries, by focusing on the tests (the transfer conditions expressed on the SL entry) and actions (the transfer conditions expressed on the TL entry) that need to be expressed to establish proper multilingual correspondences.

5.1.2 Encoding entries

In the practice of entry creation, we wanted to simulate the scenario of independently built monolingual resources that are successively linked through multilingual transfer conditions. Transfer conditions have to be expressed on the monolingual information, hence a first phase of the encoding experiment has been dedicated to the creation of monolingual entries, while multilingual correspondences have been added in a second stage. Starting from an agreed list of about 90 English lemmas, we constructed three sets of SemUs (for English, French and Italian) that we used as anchor point to build two sets of multilingual entries (multi-MILEs from English to Italian and from Italian to French).

Following the PAROLE-SIMPLE model (which provided the bootstrapping base for the experiment), each monolingual entry has been described in terms of three interlinked entities, i.e. Morphological Unit (*MU*), Syntactic Unit (*SynU*) and Semantic Unit (*SemU*), which encode respectively the morphological, syntactic and semantic relevant information. When possible, the entries encoded during the PAROLE-SIMPLE project have been re-used. In the SemUs, the various types of information available in the SIMPLE model (e.g. ontological types, examples, domain information, semantic features, semantic relations, thematic roles, selectional restrictions of the arguments, etc.) have been exploited to provide a formal characterization of the selected senses of the lexical entries. On the other hand, we focused on the necessary extensions and enrichment of the original model, especially in the perspective of the jump at the multilingual level. A particularly critical issue both at the monolingual and multilingual level is represented by the dominant role of multiword expressions and collocations. These form a kind of lexicographic “no-man’s land”, which can not be easily captured with the expressive resources of standard computational lexicons. In many circumstances, it is also difficult to organize this highly context-dependent information within the main senses articulation of each word. The border between the purely lexical idiosyncrasy and the possibility of extracting useful generalization is a very thin line, whose effective characterization is nevertheless an important demand in multilingual computational lexicography.

Basically, two types of MWE linking problems arise. Each of them can be further divided into subcategories.

MWE → MWE
MWE → word

We see that the Italian noun *colpo*, which is usually translatable with the English equivalents “blow” and “stroke” needs, in many cases, a more specific translation is, depending on the surrounding linguistic context in which this noun appears. For example, when we find *colpo* in the common context: *Colpo+ di* +INSTRUMENT ([$_{NP}$[$_{N}$colpo]] [$_{PP}$[$_{P}$di] $_{NP}$[$_{N}$X]]]), we usually translate it with *stroke,* but:

if INSTRUMENT={frusta (*whip*)} then *colpo*=*lash*
if INSTRUMENT={falce (*sickle*)} then *colpo*=*sweep*
if INSTRUMENT={testa (*head*)} then *colpo*=*header*
if INSTRUMENT={tacco (*heel*)} then *colpo*=*heel print*
etc..

In the same way, when *colpo* is followed by an adjective it can be generally translated using *blow,* but:

colpo mancino = *an underhand blow*
colpo gobbo =a stab in the back
colpo basso = a hit below the belt

MWEs can be handled either through additional monolingual lexicon entries beyond those needed for monolingual analysis, creating lists of MW-SynUs:

synU: "colpo_di_frusta"
description: colpo+[prep="di"]+[N=lex] where [lex]="frusta"
……
synU: "colpo_di_tacco"
description: colpo+[prep="di"]+[N=lex] where [lex]="tacco"
……..
synU: "colpo_mancino"
description: colpo+[adj=lex] where [lex]="mancino"
etc...

Then, in the *multi*-MILE layer we must simply record the correspondences between Italian and English SynUs:

Mult_Usyn: <colpo-di-frusta_lash>
Italian_Usyn: "colpo-di-frusta"
English_Usyn: "lash"

Mult_Usyn: <colpo-di-tacco_heel-print>
Italian_Usyn: "colpo-di-tacco"
English_Usyn: "heel_print"

Mult_Usyn: <colpo-mancino_underhand-blow>
Italian_Usyn: "colpo_mancino"
English_Usyn: "underhand_blow"

Another possible strategy for dealing with MWs is to resort to additional information just in multilingual layer, by adding or prohibiting syntactic features or positions, without touching the monolingual entries. The correspondence established only at syntactic level is the most simple and direct, but we have to specify the whole range of transfer situations - between all the different layers of lexical description - for which we have to establish links. The MILE provides us a set of explicit *lexical objects* (entities of the ER-Model) that can be used to create new syntactic slots, new arguments, to constrain semantic and syntactic information via a powerful yet simple *lego*-mechanism of tests and actions.

5.2 The Lexicographic Approach : Sense Indicators as Candidates for Transfer Conditions

A second important aspect of our work in investigating the relevant lexical conditions for multilingual transfer was the analysis of the non-formalized semantic information in the average bilingual dictionary entry: the ‘sense indicators’. This term denotes the clues which the lexicographer offers in order to differentiate a series of foreign-language equivalents of the headword, as exemplified by the material in parenthesis in the partial entry for **inhale** shown below: *breathe in* and *in smoking* respectively distinguish for the English speaker the sense of *inspirer* from that of *avaler la fumée*

> **inhale** / In"heIl /
>
> **II** *intransitive verb* (breathe in) inspirer**;** (in smoking) avaler la fumée; **to inhale deeply** inspirer profondément**.**

The focus of our analysis was the material inserted into the *Collins Gem English-French Dictionary* entries in order to guide the source language (English) speaker, encoding into French, to an appropriate target language equivalent. This type of material, often italicized and/or in parenthesis in dictionaries, is notoriously the most difficult to parse in the treatment of machine-readable dictionaries, and indeed has never before been classified[5]. Our hypothesis was that these clues, or sense indicators, if formalized, might correspond to a large extent to the transfer conditions of machine translation. We set out to extract the sense indicators semi-automatically from the dictionary text, classify them, and from this data create a pool of structured information which would allow us to evaluate these indicators as candidates for transfer conditions. Our goal was to build a database which would go some way towards defining the type of information which the human dictionary user finds useful when trying to identify translation equivalents

This section of the deliverable contains the following parts:

1. a brief account of the lexicographic process, in order to situate the use of sense indicators in the bilingual dictionary entry, together with a discussion of the frame semantics approach to lexicographic relevance;
2. a report on how our classification of the sense indicators was developed;
3. a brief description of the database in its final form;
4. a comparison of the information offered by dictionary sense indicators with the information currently used in the Comprendium MT system.

5.2.1 The lexicographic process and lexicographic relevance

It is convenient to consider the lexicographic process in two distinct phases (see Atkins (1993) for a fuller account of this). In the initial stage (*analysis*) of dictionary compiling, when lexicographers are studying the way the word behaves in the language, they look at the evidence (corpus data, their own notes etc.), record facts about the headword as they find them (meanings, constructions, collocates, participation in multiword expressions, register, language variety, style etc.), establish provisional sense distinctions, attempt to order the facts and the exemplifying sentences according to these distinctions, and thus create a rich database entry from which may be extracted the material needed for the particular dictionary they are working on. The greatest danger at this stage is that, if

[5] But see Fontenelle (1997) where he describes the use of some of this material in his conversion of the Collins-Robert English-French Dictionary into a lexical database.

there is no theoretical basis for the analysis, the collection of facts will be patchy and inconsistent, without any means of ensuring that no important aspect of the word' s behavior has been overlooked.

When editors come to the task of formulating the actual dictionary entry (*synthesis*), in the absence of any theoretical underpinning there is no means of ensuring that their approach to these tasks is consistent from A to Z of the dictionary - a process which may cover a number of years and involve a large team of lexicographers. For every entry and subentry, the major decision is what to put in – or more tantalizingly, what to leave out; obviously, native-speaker intuition informs the selection, but here again, objective facts (which normally amount to no more than frequency statistics in the current corpus) are thin on the ground. It is essential that the selection is based on a clear overview of how the word actually behaves, a good counterweight to the salient usages available to native-speaker intuition (see Hanks (2000) for a discussion of social versus cognitive salience). So much for monolingual dictionaries. In the case of bilingual dictionaries, of course, the synthesis stage also includes the 'transfer' process, when the source-language items are translated into the target language, and the entry crafted to be as helpful as possible to its eventual readers. Here, quality depends on good decisions being made at two different points in the process: first, the selection of material to go into the entry, and second, the identification in the target language of the most appropriate equivalent or set of equivalents for the headword in its various uses (see Fillmore & Atkins (2000) for a discussion of this in the context of frame semantics).

5.2.2 Analysis stage

During the analysis stage the lexicographer needs both *source* and *linguistic* information in order to judge what is relevant to the dictionary entry and what is not. Facts about the source or context from which a citation is drawn allow considered judgments about the status of the citation: whether the usage is typical of the linguistic community whose language is being recorded, and/or whether some indicator of language variety, style, register, currency and so on is required at this point. The types of source information normally available in the header of the corpus text include the title of the work, its date of publication, the genre it belongs to, and its authorship, including probably the sex, age and regional origin of the author; the header may also include other details such as regional variety, language level or domain. A different type of source information relates to collocation: the significance of the words with which the keyword combines in the corpus citations, the importance allocated to the frequency of the various senses in the corpus, and so on. (Note that frequency statistics have to be read in conjunction with the corpus design criteria before they can be evaluated.) We shall not deal with frequency in this paper.

The linguistic information relevant to the analysis stage of dictionary compiling is outlined in Fillmore & Atkins (1998) and developed in Fillmore et al. (forthcoming). In brief, this information consists of the syntactico-semantic valence of the keyword reflecting essentially the valence instantiated in the corpus.

This theory of lexicographic relevance based on frame semantics informs the current FrameNet project[6]. Although this project's thorough, exhaustive approach to corpus lexicography is impractical for the professional lexicographer working within specific time and length constraints, a simplified version of this approach has proved helpful. It is useful for lexicographers analyzing

[5] This research project, of considerable importance to professional lexicographers, is based in the International Computer Science Institute, Berkeley, California, and led by Charles J. Fillmore, whose work in frame semantics and construction grammar informs the lexicography. See http://www.icsi.berkeley.edu/~framenet/

word senses to start from the semantic valence, identify its essential components, and note the way in which these are grammatically and lexically realized in the corpus.

The teachers and medics were	arguing	about who has what of my time
This is a key factor	arguing	against the existence of such a relationship
“You’ll stop	arguing	and do as you’re damned well told!”
We spent most of our time in cafes	arguing	and holding hands
These features	argue	for a local origin.
Margaret Mead	argues	for a nurture perspective on behavior.
There was a lot of	arguing	going on between Mum and Dad.
Dr Wilson	argues	that if ants disappeared, most of …
Richard Dawkins has	argued	that it is their genes that survive.
This situation	argues	that a serious tax should be levied.
The popular press have	argued	the case.
The platoon commander was	arguing	with a gang of Christian Phalangists.

Table 4: KWIC concordances for *argue*

The verb *argue*, for which some concordances are shown in Table 4, will serve as a small case study. Scanning them, the lexicographer begins to feel her way around the word: you *argue about* something (one sense here – ‘quarrel’) but you can also *argue for* and *against* something – is that the same sense? Or a second one – ‘make a case, maintain’? Looking at the subjects of the verb in the corpus (*critic, economist, proponent, author, feminist* etc.) reinforces the two-sense view, and just as we are beginning to believe that *argue* can be described in terms of these two senses alone, we notice *this situation argues that a serious tax should be levied*, and are forced to add a third sense to our armoury, that of ‘be evidence of, indicate’, noting that in this sense the subject of *argue* is a fact, event or situation, and not a person. This third sense – rare in the 100-million-word British National Corpus - is absent in many respected dictionaries,[7] yet an ad hoc sweep of the web produces many instances like those shown in Table 5.

Cold, hard facts	argue	against the death penalty.
And do not the facts	argue	in favor of the contrary view?
And the law and the facts	argue	strongly for continuation of the lawsuit.
…though the statistics	argue	otherwise:
Women' s Caucus statistics	argue	that unfairness does not happen to men …
The resulting statistics	argue	that they have orbits with inclinations near 23°.

Table 5: From the web, *argue* in the ‘indicate’ sense

The lexicographers would have been less likely to overlook that third sense had they taken a FrameNet approach to the analysis of the corpus data. This involves identifying the *frame*, or conceptual background, to which the lexical unit[8] belongs, then discovering how the various *elements* in the frame are realized in the corpus sentences. In the case of *argue* (and *argument*) the three relevant frames are:

1. Communication-Conversation (e.g. *She knew better than to argue with him.*)
2. Communication-Arguing (e.g. *He argued that it was unconvincing.*)
3. Reasoning-Evidence (e.g. *Cold hard facts argue against the death penalty.*)

It is clear that this third sense does not belong to a communication frame, which is described in terms of the frame elements INTERLOCUTOR-1, INTERLOCUTOR-2 and TOPIC, *inter alia*.

[7] For instance, Cobuild English Dictionary, Oxford New Dictionary of English, Concise Oxford Dictionary, among others.

[8] The equivalent of the word in one of its senses.

It is impossible to do justice to the complexity of the FrameNet approach[9] in the space available to us here, but, taking the ‘quarrel’ sense of *argue* and *argument* as an example, we shall attempt to show how a cut-down version of this can help day-to-day lexicographic analysis. Figure 3 shows a composed example sentence for the verb, analyzed in this way.

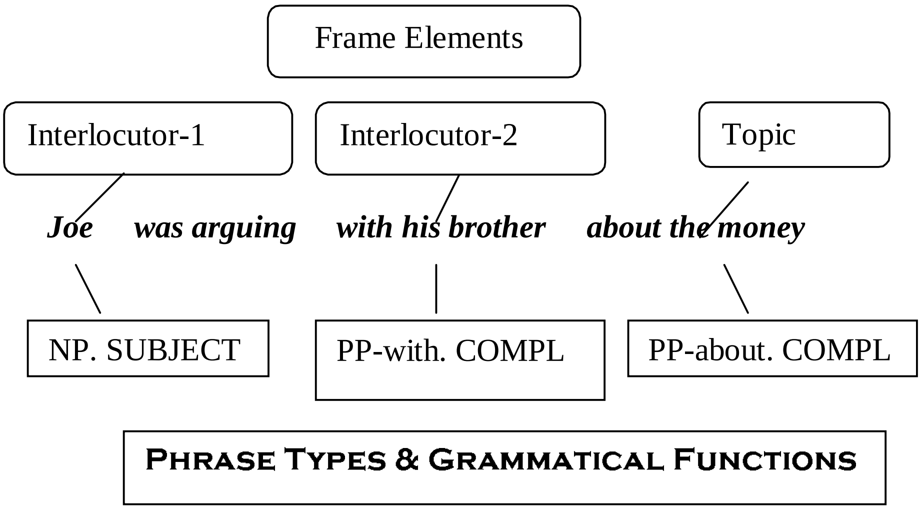


Figure 3: Frame analysis of ‘quarrel’ example

This lexical unit belongs to the Conversation frame, of which three core elements[10] are instantiated in this sentence, as follows:

INTERLOCUTOR-1 : one of the parties involved in the conversation;
INTERLOCUTOR-2 : the other party involved
TOPIC : the subject of conversation – in this case, what they are quarrelling about.

These core elements encapsulate the essential grammatical facts which this sentence offers to the lexicographer, identifying the following ‘chunks’ of the sentence as lexicographically relevant:

Joe : a noun phrase (NP) functioning as the SUBJECT of the keyword *argue*;
with his brother : a prepositional phrase (PP) functioning as the COMPLEMENT of the keyword *argue*; and
about the money : another prepositional phrase (PP) functioning as the COMPLEMENT of the keyword *argue*

In FrameNet terms, the frame elements in this sentence and their grammatical instantiation constitute a ‘valence pattern’ for the lexical unit *argue* in the Conversation frame. The set of valence patterns identified in an exhaustive search of corpus data constitutes the valence of this

[9] A full account of this project will be given by the FrameNet team in a forthcoming edition of the International Journal of Lexicography, to be guest-edited by Thierry Fontenelle and scheduled for 2003.

[10] A "core" versus "periphery" distinction is established among the frame elements that accompany a frame-bearing word, the former indicating those that are most closely associated with the meaning of the headword, the latter covering expressions of time, place, manner, etc., that provide modifications of the sort that could be added to almost any situation type. Core frame elements include: obligatory objects and complements of the headword; any frame element which, if expressed, would be expressed as direct object of a verb headword, or as a PP-of in the case of the corresponding noun; any frame element which, if unexpressed, is interpreted as a case of definite null instantiation (such as the thing you are blaming John for when you say I blame John).

lexical unit. In terms of the needs of the professional lexicographer, the valence identifies all of the facts needed for a full description of the word's corpus behavior (apart from frequency and collocational data).

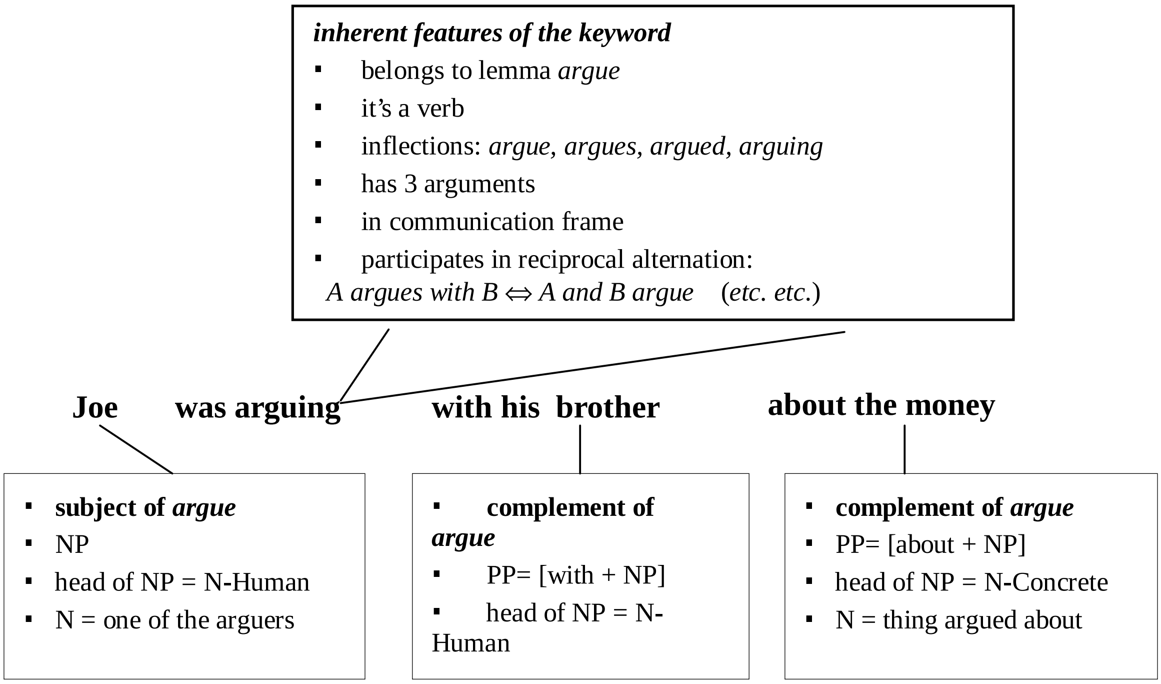


Figure 4: Inherent and contextual features of *argue*

Let us look for a moment at what the single sentence *Joe was arguing with his brother about the money* tells us about *argue*, namely the information shown in Figure 4 as the 'contextual features of the keyword'. We know from this one sentence that *argue* can occur in a continuous tense; that it can be used with two complements; and that the verb's subjects and its complements can be instantiated by words with the properties detailed in the diagram. All of that – together with the 'inherent features' also shown – constitutes lexicographically relevant information. These are facts about the word which must be taken into account by anyone writing a dictionary entry for *argue*, and if the dictionary is destined for encoding language learners, and is of any reasonable size, the contextual features must figure in the entry. If they are not there, the language learner cannot use the word correctly. If they are not in a bilingual entry, together with their target-language equivalents, the verb's full potential cannot be expressed in the foreign language.

Figure 5 needs no detailed commentary: we include it to demonstrate the amount of lexicographically relevant information which a single corpus sentence offers about nouns, which tend to be second-class citizens in the world of lexicography.

In summary, the lexicographically relevant information for each word sense, needed by the dictionary editor during the analysis stage of the process, may be described as:

- the inherent features of the keyword itself (e.g. it's a verb, etc.)

- and the various details of its contextual features.

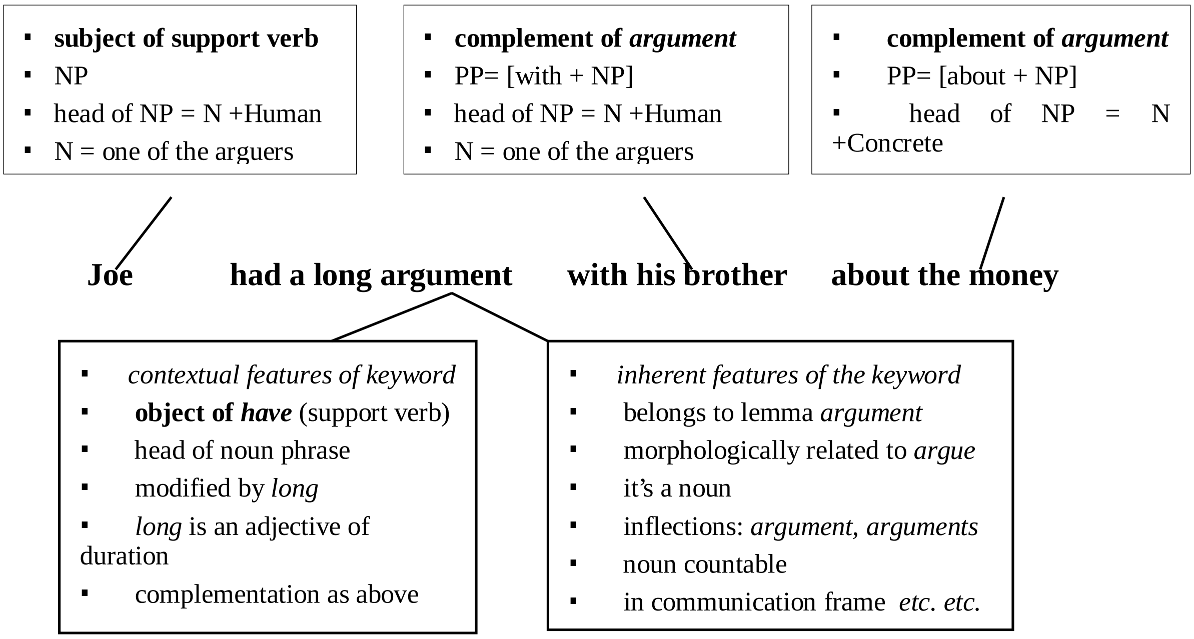


Figure 5: Inherent and contextual features of the *argument*

These features are, for each frame element expressed in each of the varied corpus contexts:

- its semantic role (e.g. one of the arguers)
- its grammatical function (e.g. subject of *argue*)
- its phrase type (e.g. NP)
- the sortal feature of the head noun of the NP (e.g. 'human').

5.2.2.1 SYNTHESIS STAGE

For editors of bilingual, as opposed to monolingual dictionaries, the synthesis stage (the extraction of the dictionary entry from the monolingual database entry) is complicated by the fact that it must also contain the 'transfer' process, whereby target language equivalents are proposed and evaluated, and selected or rejected, and the structure and content of the entry is subject to changes motivated by the needs of the users. If the dictionary is destined for use by speakers of both the source and the target languages, then its editors must keep this fact at the forefront of their minds throughout the work of compiling the entry. Indeed, if this is the case, then the dictionary must be two dictionaries rolled into one (and, as such, it will inevitably contain some redundant information for both sets of users). An example of this is given in Figure 6, where alternative entries for the

French noun **couche** show how the needs of the encoding Francophone override those of the decoding Anglophone. Such a dictionary entry is bound to favor the source-language speakers, who need much more help and guidance in formulating sentences in a foreign language than do the target-language speakers. The latter are simply trying to understand an expression in the foreign language, and sometimes to find its equivalent in their own. They are unlikely to select an item in their own language which is manifestly at odds with its context.

A

couche[1]
nf (*pour bébés*) nappy (*Brit*), diaper (Am).
couche[2]
nf **1** (*de vernis, peinture, d'apprêt*) coat; (*d'aliments, de poussière*) layer. **2** (*strate*) stratum, layer.

B

couche[1]
nf nappy, diaper.
couche[2]
nf coat; layer; stratum.

Figure 6: Entries for the francophone (A) and anglophone (B)

The entries in Figure 6 marked (A) would appear in a dictionary prepared for both French and English markets; the entries marked (B) show how much of that information the English speakers really need. They know when it is appropriate to select *nappy* or *diaper,* and don't need to be told that one is British and one American English. They know when *nappy* is an appropriate choice for their English context, and when they should prefer *coat* or *layer* or *stratum*. The French speaker has to be guided to the appropriate English word, by *pour bébés* ('for babies'), or *de vernis, peinture, d'apprêt* ('of varnish, paint, size') or *d'aliments, de poussière* ('of food, dust') or *strate* ('stratum').

5.2.2 Classifying the Sense Indicators in the Database

develop / dI"vel@p / *vi*
(*evolve*) *child, seed, embryo* se développer; *intelligence* s' épanouir; *skills* s' améliorer;*society, country*

act / &kt / *n*
1 (*action, deed*) acte *m.* 2 *Law, Politics* loi *f*; **Act of Parliament/ Congress** loi votée par le Parlement/le Congrès; …

Figure 7: Three types of sense indicators

When it comes to compiling the actual entry, the lexicographer opens a dialogue with the dictionary user, and relies on different types of facts in order to help the user understand the entry. In the case of a bilingual dictionary, the entry is rich in indicators (like the italicized material in Figure 7) whose function is to guide the reader to the appropriate foreign-language expression. A detailed listing of the types of lexicographically relevant information used in the FrameNet analysis formed the starting point of our classification of sense indicators. However, it proved inadequate to classify comprehensively the richness of material extracted from a small pocket English-French dictionary, and the classificatory system and metalanguage had to be greatly expanded.

We began by classifying the indicators into four different types. In the Figure 7, the synonyms *evolve* in the **develop** entry and *action, deed* in the **act** entry belonged to a class of **hierarchical indicators**, which, as well as synonymy, included instances of antonymy, hyperonymy, hyponymy and meronymy. The various typical subjects of the verb *develop* (*child, seed, embryo; intelligence; skills; society, country* etc.) belonged to the class of **morphosyntactic indicators**, while the *act* entry offered an example of **language sub-type indicators** (*Law, Politics* indicate domain: in legal and political contexts the equivalent of English *act* is French *loi).* We identified a fourth class of sense indicators which we call **semantico-syntactic indicators**: these are exemplified in Figure 8 by the various synonyms of *set* (*collection, kit, game, pair, group, scenery* etc.) used to clarify the sense distinctions and lead the Anglophone reader to the appropriate translations.

set /set / **I** *n*
1 (*collection*) (*of keys etc.*) **jeu** ***m*; …**
2 (*kit, game*) **a chess set** un jeu d' échecs;
3 (*pair*) **a set of sheets** une paire de draps; **…**
4 *Sport* (*in tennis*) …
5 (*television*) …
6 (*group*) (*social*) monde *m*; (*sports*) milieu *m*; …
7 (*scenery*) theatre décor *m*; …
9 *GB school* (*class, group*) groupe *m*; …
10 (*hair-do*) mise *f* en plis; …
11 *music* concert ***m*;**
12 (*position*) (*of sails*) réglage ***m*;** …
13 (*direction*) (*of wind*) sens ***m*;** … *etc. etc. etc.*

Figure 8: Synonyms as sense indicators

The material which constitutes the database was semi-automatically acquired from the *Collins Gem English-French Dictionary.* The acquisition of the data was done in three stages.

1. In the first stage, all the headwords were automatically extracted from the dictionary, with their translations and the corresponding sense indicator.
2. The sense indicators were then manually classified with regards to the lexicographically relevant features (LRFs) derived from the FrameNet analysis as defined above, i.e. hierarchical, syntactic, semantico-syntactic and language subtype.
3. This material was taken as input to generate an SQL database.

Thus, for instance, the **develop** entry shown in Figure 9[11] was converted into the representation given in Figure 10, where the various fields indicate respectively the headword (*develop*), the indicator part of speech (*N*, for noun, *Abbr* for abbreviation, *V* for verb in this example), the indicator itself (*gen, habit*, etc.), the type of sense indicator (for *Syn* for synonymy; *Subjv*, for

[11] This is the actual entry in the Collins Gem English-French Dictionary, which populates the database; other entries used as illustrative material in this paper are modelled on those in larger English-French dictionaries, such as the Collins-Robert and the Oxford-Hachette.

subject of the verb; etc.), the headword part of speech (*vi* for intransitive verb and *vt* for transitive verb), the translation of the headword and, eventually, the context (obligatory preposition, etc.).

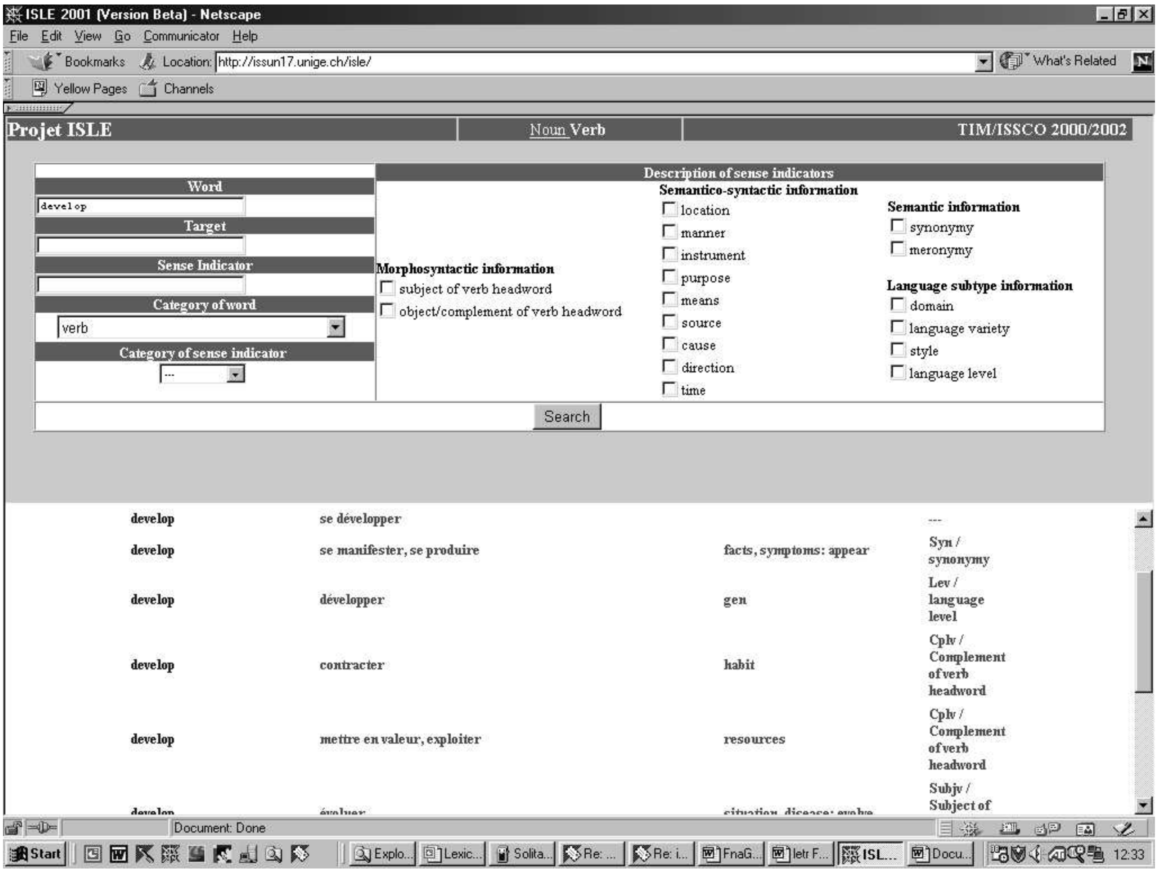
develop [dI"vel@p]) *vt (gen)* développer; *(habit)* contracter; *(ressources)* mettre en valeur, exploiter *vi* se développer; (*situation*, *disease*; *evolve*) évoluer; (*facts*, *symptoms*: *appear*) se manifester, se produire.

Figure 9: Collins Gem entry for *develop*

```
Headword¦indicator p.o.s¦indicator¦headword  p.o.s¦translation¦context
develop|Abbr|gen|Lev|vt|développer¦
develop|N|habit|Objv|vt|contracter¦
develop|N|resources|Objv|vt|mettre en valeur, exploiter¦
develop||||vi|se développer¦
develop|N|situation, disease; evolve|Subjv|vi|évoluer¦
develop|N|symptoms: appear|Subjv|vi|se manifester, se produire¦
develop|V|facts, symptoms: appear|Syn|vi|se manifester, se produire¦
develop|V|situation, disease; evolve|Syn|vi|évoluer¦
```

Figure 10: Representation of the information

Figure 11 shows the early format of the database query screen, which was later considerably simplified to give the final version. The top part of the screen is the query interface--here the user searches for the word *develop*. The bottom part shows the result of the query, i.e. the headword in the first column, then the different translations, the sense indicators and its type as classified in the database.



Location: http://issun17.unige.ch/isle/

Projet ISLE | Noun Verb | TIM/ISSCO 2000/2002

Word: develop

Target:

Sense Indicator:

Category of word: verb

Category of sense indicator: ---

Description of sense indicators

Morphosyntactic information
□ subject of verb headword
□ object/complement of verb headword

Semantico-syntactic information
□ location
□ manner
□ instrument
□ purpose
□ means
□ source
□ cause
□ direction
□ time

Semantic information
□ synonymy
□ meronymy

Language subtype information
□ domain
□ language variety
□ style
□ language level

develop	se développer		---
develop	se manifester, se produire	facts, symptoms: appear	Syn / synonymy
develop	développer	gen	Lev / language level
develop	contracter	habit	Cplv / Complement of verb headword
develop	mettre en valeur, exploiter	resources	Cplv / Complement of verb headword
develop	évoluer	situation, disease: evolve	Subjv / Subject of

Figure 11: Original query screen

As can be seen from Figure 11, the tool offers the possibility not only of searching for a headword, a target word, or a particular sense indicator (e.g. *person*, CULIN, etc.), but also of extracting all indicators that have a specific syntactic structure (*with N*, *by N* etc.) or belong to a particular type (*Loc*, *Man* etc.), according to the classification discussed above.

The interim version of the database which supported this query programme consisted of four separate tables, given here as Table 6, 7, 8 and 9.

Head word	**Type**	**Description of indicator**	**Headword**	**Sense Indicator**	**TL equivalent**
verb	Subjv	indicator is subject of verb headword	develop	situation, disease	évoluer
verb	Objv	indicator is object of verb headword	develop develop	habit resources	contracter mettre en valeur, exploiter
noun	Subjn	indicator is subject of verb base of noun headword	contortion	of acrobat	contortion
noun	Argn	indicator is argument of support verb of noun headword	pulse	of heart	battement

noun	Possn	indicator is possessor of noun headword	web	of spider	toile
noun	Adj-Pertn	indicator is pertainym modifying noun headword	bell	electric	sonnerie
noun	Gen	indicator is gender of headword	mousse	feminine masculine	moss cabin boy
adjective	Moda	indicator is noun typically modified by headword	vivid	account	frappant(e)

Table 6: Morphosyntactic information in the original database

Headword	Type	Description of indicator	Headword	**Sense Indicator**	**TL equivalent**
noun / verb	Loc	indicator is location (typical environment) of headword	promenade	by sea	esplanade, promenade
noun / verb / adjective	Man	indicator is manner of headword	inferior	in rank	subalterne
noun / verb	Inst	indicator is instrument of headword	sign	with hand etc	signe
noun / verb/ adjective	Pur	indicator is purpose of headword	bat	for baseball etc	batte
noun / verb	Means	indicator is means of headword	passage	by boat	traversée
noun / verb	Sou	indicator is source of headword	deduction	from wage etc	prélèvement, retenue
noun / verb	Tim	indicator is time of event of headword	blackout	in wartime	couvre-feu
noun / verb	Dir	indicator is direction of headword	move	forward	avancer
noun / verb / adjective	Cau	indicator is cause of headword	hangover	after drinking	gueule de bois
noun	Coll	indicator is items collected by collective headword n	block shift	of buildings of workers	pâté (de maisons) équipe
noun	Mass	mass n indicator (of itemiser noun headword)	pinch sheet	of salt etc of paper	pincée feuille
noun	Conte	indicator is contents of headword noun container	can	of milk, oil, water	bidon
noun	Conta	indicator is container of headword noun contents	arrow	in quiver	flèche
noun	Des	indicator describes noun headword	bolt lodger ferry	with nut with room & meals small	boulon pensionnaire bac
noun	Mat	indicator is material of which headword is made	moulding	in wood	moulure
noun	Cpln	indicator is complement of	gallantry	towards	galanterie

		noun headword		ladies	
noun	Cplna	indicator is complement of an agentive noun headword	trainer	of dogs etc.	dresseur/ euse
verb	Coplv	indicator is complement of verb headword	protect	against attack	protéger
adjective	Cpla	indicator is complement of adjective headword	gallant	toward ladies	empressé, galant

Table 7: Semantico-syntactic information in the original database

Type	Description of indicator	Headword	**Sense Indicator**	**TL equivalent**
Syn	indicator is synonym of headword	casual stress	by chance accent	fortuit accent
Ant	indicator is 'not' + antonym of headword	wrong	not suitable	qui ne convient pas
Hyper	indicator is hyperonym of headword	spinach flying	food activity	épinards aviation
Hypo	indicator is hyponym of headword			
Mer	indicator is the whole; headword is the part	stone butt	in fruit of cigarette	noyau mégot

Table 8: Hierarchical information in the original database

Type	Description of indicator	Headword	**Sense Indicator**	**TL equivalent**
Dom	indicator is domain of headword, e.g. Architecture, Music	grant	Admin	subside, subvention
Lgv	indicator is language variety of headword, e.g. American/British	automobile	US	automobile
Sty	indicator is style of headword, e.g. informal, jargon	aim	fig	viser (à)
Lev	indicator is language level of headword, e.g. general language, technical language, etc.	canvas	gen	toile

Table 9: Language subtype information in the original database

5.3 A database of sense indicators

The classification in Table 6-9 proved unsatisfactory, however. It turned out to be impossible to define objective criteria for the assignment of categories to all the indicators, and we could not adduce a theoretical basis for many of the decisions we made during the classification. We therefore decided to simplify the database, holding it in dual format which distinguishes between 'contextual' information (i.e. patterns which may be found in the context of an actual lexical item in a text for translation), and 'inferential' information (knowledge intended to be inferred from the indicators). In this simplified version, many of the arbitrarily assigned classes in Table 6 and particularly Table 7 – such as Location, Manner, Purpose, Means, Source etc. – are simply described as ‘Inferred Domain’, since we believe that the reason why the lexicographer offered this information was to allow the user to infer the domain, or subject matter, or in broader terms the type of context, which applied in the case of the equivalent being designated as appropriate for the indicator in question.

The simplified database may be consulted at http://issun17.unige.ch/isle, and has the structure described in the tables in Tables 10 and 11. These show the different types of sense indicator which we felt confident about identifying for each category found in the Collins Gem dictionary. However, since this very small work uses only a subset of sense indicators which are standard for larger English-French dictionaries in the Collins series, the table also includes some of the more common indicators missing from the Gem dictionary. These additions are not included in the online database. In Figure 16, the indicators giving ‘contextual’ information stand in specific syntactic relationships to the headword and these are shown the **Indicator Type** column.

Headword P-O-S	Indicator Type	Description of Indicator	Example		
			Source-Language Headword	Sense Indicator	Target-Language Equivalent
verb	Subjv	indicator is subject of verb headword	develop	situation, disease	évoluer
verb	Objv	indicator is object of verb headword	develop develop	habit resources	contracter mettre en valeur, exploiter
noun	Argn	indicator noun is argument of verb base of noun headword	acting	of actor	jeu
noun	Moda	indicator is adjective typically modifying headword	bell	electric	sonnerie
noun, adjective	Modn	indicator is noun typically modified by headword	vivid	account	frappant(e)
all p.o.s	ofPP[1], atPP, fromPP, toPP, etc.	indicator is a PP that modifies the headword	escape	from jail	s'évader

Table 10: Contextual information in the Sense Indicators Database

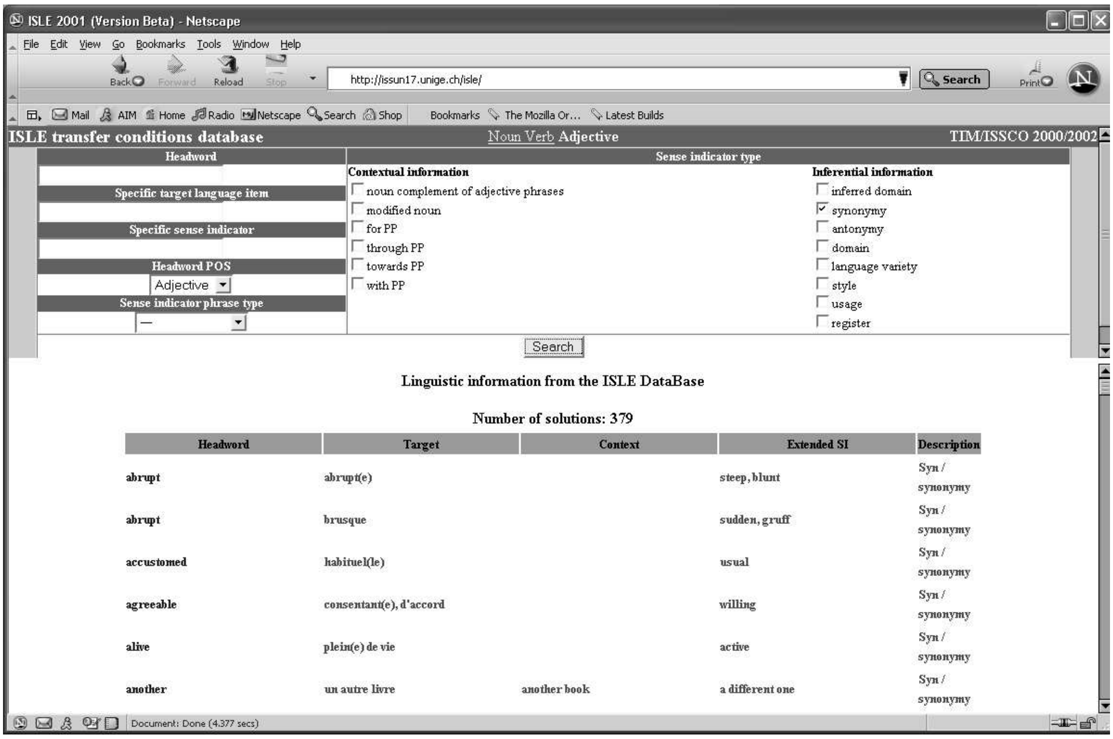
[1] This type is realized in actual text as (for instance) either *the family's wealth* or *the wealth of the family*.

Headword P-O-S	Indicator Type	Description of Indicator	Example		
			Source-Language Headword	Sense Indicator	Target-Language Equivalent
all p-o-s	Ant	indicator is ‘not’ + antonym of headword	wrong	not suitable	qui ne convient pas
all p-o-s	Syn	indicator is broad synonym of headword	spinach flying	food activity	épinards aviation
all p-o-s	Dom[1]	indicator denotes domain of headword, e.g. Architecture, Music	grant	Admin	subside, subvention
all p-o-s	Lgv[1]	indicator denotes language variety of headword, e.g. American/British	automobile	US	automobile
all p-o-s	Sty[1]	indicator denotes style of headword, e.g. informal, jargon	aim	fig	viser (à)
all p-o-s	Lev[1]	indicator denotes language level of headword, e.g. general language, technical language, etc.	canvas	gen.	toile
all p-o-s	Reg[1]	indicator denotes the register of the headword	bamboozle	col	embobiner
all p-o-s	Inferred domain	indicator is an inferred domain	promenade	by sea	esplanade, promenade

Table 11: Inferential information in the Sense Indicators Database

[1] Indicators of this type are normally drawn from a set defined extensively in the forematter or backmatter of the dictionary.

The pared-down database gave rise to a less complex query screen. The final version of the GUI for adjective queries is shown in Figure 12.


ISLE 2001 (Version Beta) - Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://issun17.unige.ch/isle/
Search
Print
Mail AIM Home Radio Netscape Search Shop
Bookmarks The Mozilla Or... Latest Builds
Document: Done (4.377 secs)

ISLE transfer conditions database | Noun Verb Adjective | TIM/ISSCO 2000/2002

Headword

Specific target language item

Specific sense indicator

Headword POS: Adjective

Sense indicator phrase type: —

Sense indicator type

Contextual information
- □ noun complement of adjective phrases
- □ modified noun
- □ for PP
- □ through PP
- □ towards PP
- □ with PP

Inferential information
- □ inferred domain
- ☑ synonymy
- □ antonymy
- □ domain
- □ language variety
- □ style
- □ usage
- □ register

Search

Linguistic information from the ISLE DataBase

Number of solutions: 379

Headword	Target	Context	Extended SI	Description
abrupt	abrupt(e)		steep, blunt	Syn / synonymy
abrupt	brusque		sudden, gruff	Syn / synonymy
accustomed	habituel(le)		usual	Syn / synonymy
agreeable	consentant(e), d'accord		willing	Syn / synonymy
alive	plein(e) de vie		active	Syn / synonymy
another	un autre livre	another book	a different one	Syn / synonymy

Figure 12: the adjective query screen – final version

This query interface is self-explanatory for the most part. Suffice it to say that one or more of the dialog boxes may be activated for any query; naturally, the more dialog boxes activated, the more specific the query, and the fewer the results returned. In the screen in Figure 12, the search specified ‘all indicators of the type ‘synonymy’’, i.e. all indicators which were synonyms or pseudo-synonyms of the headword. We note in the results that the entry for the adjective **abrupt** contains two indicator groups responding to the search conditions: the first (*steep*, *blunt*) pointing to the French *abrupt(e)* as appropriate in the context of, for instance, *abrupt drop*, or *abrupt cliff*, while the second (*sudden*, *gruff*) points to *brusque* for contexts such as *abrupt answer* or *abrupt manner*. The complete entry from the Gem dictionary is shown in Figure 13.

abrupt [@"brVpt] *a* (*steep*, *blunt*) abrupt(e); (*sudden*, *gruff*) brusque.

Figure 13: : *abrupt* entry from Gem Dictionary

The database contains all of the sense indicators extracted automatically from the Gem Dictionary; the classification attempts to make explicit the various types of sense-indicating, non-parsable information which lexicographers use in the compilation of a bilingual dictionary in order to guide the dictionary user to the appropriate translations. As well as contributing to further granularity in the analysis of semantic information in the lexical entry, the purpose of the Sense Indicator Database is to lead to a finer parse of these dictionaries and hence to identifying more detailed transfer conditions for machine-translation systems. On this basis, it is possible to compare the sense-indicating information in the Gem dictionary with the transfer conditions in a real-world MT system, and a few examples of such a comparison are included in the next section.

5.2.4 Comparison with MT dictionaries : the case of Comprendium

The distinction between contextual and inferential information corresponds to the distinction between information that can be imported today into a MT system and information that requires manual intervention before this is possible. However, when we compare our database with the Comprendium[12] bilingual dictionary, it is striking to see that even small traditional dictionaries like the *Gem* make specific more possible translations than the MT dictionary. In order to illustrate this, some comparisons between a few entries in the Comprendium and the same entries in the Gem dictionaries are shown in Table 12-19. Note that the sign "----" is used to indicate that the translation is not present in that particular dictionary.

Entry	**rare**	**Comprendium**	**Gem dictionary**
Translations	rare	default	default
	raréfié	-----	mod=air/atmosphere

	saignant	mod=head/steak	mod=steak,
			DOMAIN=CULIN

Table 12: the adjective *rare*

Entry	**rocky**	**Comprendium**	**GEM**
Translations	précaire	default	-----
	rocheux	----	mod=hill
	rocailleux	----	mod=path
	branlant	----	mod=unsteady:table

Table 13: the adjective *rocky*

Entry	**ball**	**Comprendium**	**GEM**
Translations	balle	default	for tennis, golf
	boule	-----	default
	bal	-----	dance
	ballon	-----	football

Table 14: the noun *ball*

[12] Comprendium is a commercial transfer-based system, developed by Sail Labs (http://www.sail-labs.de/). The data used here is drawn from the default package of the system.

Entry	**cut**	**Comprendium**	**GEM**
Translations	coupure	default	default
	reduction	prep=in	in salary
	coupe	-----	of clothes
	taille	-----	of jewel
	morceau	prep=of	of meat

Table 15: the noun *cut*

Entry	**nut**	**Comprendium**	**GEM**
Translations	noisette	default	default
	écrou	----	of metal

Table 16: the noun *nut*

Entry	**removal**	**Comprendium**	**GEM**
Translations	enlèvement	default	taking away
	déménagement	----	from house
	ablation	----	MED
	renvoi	----	from office: sacking
	suppression	----	taking away

Table 17: from the entries for the noun *removal*

Entry	**walk**	**Comprendium**	**GEM**
Translations	accompagner	default	----
	marcher	no direct object	default
	marcher	direct object = tyn/loc/unit	----
	faire à pied	---	distance
	promener	direct object=ani	dog
	se promener	----	for pleasure, exercise

Table 18: the verb *walk*

Entry	**scratch**	**Comprendium**	**GEM**
Translations	grater	no direct object = a-sem/abs	default
	se grater	no direct object	default
	griffonner	direct object = a-sem/abs	---
	érafler	----	paint, etc.
	rayer	----	record
	griffer	----	with claw, nail

Table 19: the verb *scratch*

The distinction between contextual and inferential information is critical when it comes to importing these new translations into the MT system. **Contextual sense indicators** can be directly imported. As we saw in the examples, Comprendium allows local tests on the canonical form in subject position, object position, attached PP, modified noun and the descriptive adjective. The methodology would involve:

(a) checking if the translation is available;

(b) if not, translating the Gem sense indicator into the corresponding Comprendium transfer condition.

In the current version of the system, tests on PP for verbs are possible only if the PP is attached to the verb in the syntactic representation. It is therefore possible that some of the tests involving PP will not be possible in the current version of the system because the syntactic tree is not rich enough.

In the case of **inferential sense indicators**, the corresponding translation is generally not included in the MT dictionary, or else constitutes the default translation, as shown in Table 20 and 21. These entries deal with nouns whose SI is an inferred domain (*withPP* in Table 20 and *inPP* or *Adv* in Table 21). The methodology in these cases would involve inferring non-automatically the domain that is implied by the SI, or listing the individual contexts in which this translation is possible in the meaning described.

Gem dictionary		
rub	with cloth	coup de chiffon ou de torchon
	on person	friction
sign	default	signe
	with hand	signe, geste
	notice	panneau, écriteau
stab	with knive	coup de couteau

Comprendium		
rub	default	frottement
sign	default	signe
stab	default	coup de couteau

Table 20: Inferred domain (withPP) – Example 1

Gem dictionary		
extravagant	default	extravagant
	in spending	prodigue, dépensier
promiscuous	sexually	de moeurs légères
sunburnt	default	bronzé, hallé
	painfully	brûlé par le soleil
top	default	du haut
	best	meilleur
	in rank	premier

Comprendium		
extravagant	default	dépensier
promiscuous	default	immoral
sunburnt	----	----
top	mod=wine/choice/restaurant/buy	bon
	mod=job/profession	élevé
	mod=HUM/SOC	grand
	default	premier
	mod=LOC/BPART	supérieur

Table 21: Inferred domain (inPP or Adv) – Example 2

6 The ISLE Recommendations. The Multilingual ISLE Lexical Entry (MILE)

With this chapter we start the "recommendations" part of the Guidelines, based on all what was described in the previous sections.

6.1 Basic EAGLES principles

We remind here just a few basic methodological principles derived from and applied in previous EAGLES phases. They have proven useful in the process of reaching consensual *de facto* standards in a bottom-up approach and are at the basis also of ISLE work.

The MILE is envisaged as a highly ***modular*** and ***layered*** structure, with different levels of recommendations. Such an architecture has been proven useful in previous EAGLES work, e.g in the EAGLES morphosyntactic recommendations (Monachini and Calzolari, 1996), which embody three levels of linguistic information: obligatory, recommended and optional (optional splits furthermore into language independent and language dependent). This modularity would enhance: the flexibility of the representation, the easiness of customization and integration of existing resources (developed under different theoretical frameworks or for different applications), the usability by different systems which are in need of different portions of the encoded data, the compliance with the proposed standards also of partially instantiated entries. *It also provides for accommodation of very simple data types while allowing for much richer and complex models.*

The MILE recommendations are also very ***granular***, in the sense of reaching a maximal decomposition into the minimal basic information units that reflect the phenomena we are dealing with. This principle was previously recommended and used to allow easier reusability or mappability into different theoretical or system approaches (Heid and McNaught, 1991): small units can be assembled, in different frameworks, according to different (theory/application dependent) generalization principles. Such basic notions must be established before considering any system-specific generalizations, otherwise our work may be too conditioned by system-specific approaches. For example, 'synonymy' can be taken as a basic notion; however, the notion of 'synset' is a generalization, closely associated with the WordNet approach. 'Qualia relations' are another example of a generalization, whereas 'semantic relation' is a basic notion. Modularity is also a means to achieve better granularity. High granularity and maximal decomposition does not mean that we limit our recommendations to these very basic notions. On the contrary, whenever has been found consensus enough on a more complex linguistic object, we *have also provided within MILE the definition of such shareable commonly agreed linguistic objects* (e.g. synsets and qualia relations).

On the other side, past EAGLES experience has shown it is 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.), in order to i) allow for agreement on a minimal level of specificity especially in cases where we cannot reach wider agreement, and/or ii) enable mappability 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, e.g. it seemed unrealistic to agree on a set of grammatical functions. This led to an underspecified recommendation, but nevertheless one that was useful. The same possibility of underspecified (under many respects) information is provided by MILE, through a hierarchical organization of basic notions.

The principle guiding the elicitation and proposal of MILE basic notions needed to describe the multilingual level has been, according to a previous EAGLES methodology, the so-called *'edited union'* (term put forward by Gerald Gazdar in earlier EAGLES work) of what exists in major lexicons/models/dictionaries/standards, at least as a starting point, enriched with those types of information which are usually not handled, e.g. those of collocational/syntagmatic nature, and obviously those pertinent to the multilingual layer. The work of gathering descriptions and characterizations of multilingual lexical phenomena from a set of major existing lexicons, systems, dictionaries, etc., provides better ground to decide what is needed, what can be agreed on, what can be integrated in a unitary MILE, what is lacking or needs formalization, and so on. Connected to this, it is expected that any MILE proposal may contain *redundancy*. This is not problematic with regard to recommendations. It is only at the level of the specific lexicon instance that a lexicon builder may want to avoid redundancy, for reasons of efficiency, etc.

This method of work has proven useful in the process of reaching consensual *de facto* standards in a bottom-up approach and is at the basis also of ISLE work. There is every interest in building on existing resources, rather than starting from scratch, thus efforts must continue in this direction.

6.2 MILE 'basic notions'

6.2.1 Common Basic Notions and Lexical Semantics

Natural language meaning has always been thought of as one of the hardest problems for standardization. However, the increasing use of conceptual classification in the development of language technologies is rapidly changing this perception. At the same time, the growing need for dealing with semantics and contents in HLT applications is pushing towards more powerful and robust semantic components. Within the last decade, the availability of robust tools for language analysis has provided an opportunity for using semantic information to improve the performance of applications such as Machine Translation, Information Retrieval, Information Extraction and Summarization. As this trend consolidates, the need of a protocol which helps normalize and structure the semantic information needed for the creation of reusable lexical resources within the applications of focus, and in a multilingual context, becomes more pressing. Times are thus mature to start tackling the question of how to formulate guidelines for multilingual lexical (semantic) standards.

Sense distinctions are especially important for multilingual lexicons, since it is at this level that cross-language links need to be established. The same is true of syntagmatic/collocational/ contextual information. To these areas we have paid particular attention in the ISLE recommendation phase, and we have examined how to extend the available EAGLES guidelines in these and other areas to propose a broad format for multilingual lexical entries which is of general utility to the community.

In the previous EAGLES work on Lexicon Semantics (Sanfilippo et al., 1999) the following technologies were surveyed to determine which types of semantic information were most relevant:

- Machine Translation (MT)
- Information Extraction (IE)

- Information Retrieval (IR)
- Summarizations (SUM)
- Natural Language Generation (Gen)
- Word Clustering (Word Clust)
- Multiword Recognition + Extraction (MWR)
- Word Sense Disambiguation (WSD)
- Proper Noun Recognition (PNR)
- Parsing (Par)
- Coreference (Coref)

The results of the previous EAGLES survey are here summarized. Each different type of semantic information is followed by the application type in which it figures:

- BASE CONCEPTS, HYPONYMY, SYNONYMY: all applications and enabling technologies
- SEMANTIC FRAMES: MT, IR, IE, & Gen, Par, MWR, WSD, Coref
- COOCCURRENCE RELATIONS: MT, Gen, Word Clust, WSD, Par
- MERONYMY: MT, IR, IE & Gen, PNR
- ANTONYMY: Gen, Word Clust, WSD
- SUBJECT DOMAIN: MT, SUM, Gen, MWR, WSD
- ACTIONALITY: MT, IE, Gen, Par
- QUANTIFICATION: MT, Gen, Coref

It is important to notice that all of these semantic information types (except for quantification) are covered by the SIMPLE model (see above section 4.2.1). For this reason, the structure and the characteristics of SIMPLE (as a lexical resource designed on the basis of the EAGLES recommendations) has a crucial place in the design of the MILE. Within the MILE we have complemented the SIMPLE design and basic notions also with WordNet-style lexicons, thereby trying to get at a more comprehensive and coherent architecture for the development of semantic lexical resources.

Obviously ***MILE also includes previous EAGLES recommendations for other layers***. We have evaluated the usefulness of these other layers in the multilingual perspective, e.g. for the MT and CLIR tasks. We therefore had to analyze whether existing EAGLES recommendations, or existing lexicon models, with respect to the agreed basic notions, comply with the requirements of a multilingual perspective. It has however appeared that existing models (or even the union of them) do not cover all the notions/data which are needed for multilingual tasks. In this respect, we had to discover areas of deficiency, and highlight areas in need of further analysis. The same is true of applications: for some of the already available lexical information, current systems are not yet able to use it. Here too areas where systems could be easily improved could be spotted and put forward.

6.2.2 Basic Notions: operative definitions and background

Identifying the basic notions of the MILE means to understand which are the lexical dimensions that play a role, at any level of linguistic description, in a multilingual framework. The work of the previous EAGLES (focussed on Morphosyntax, Subcategorization and Lexical Semantics) provided an analysis of the linguistic information crucial for the description of a computational lexical entry in monolingual perspective. The ISLE intention was to exploit the EAGLES bulk of work and to extend the results in a multilingual perspective, trying to make a synthesis of all the information that is relevant to build a multilingual lexical entry (a MILE) starting from a monolingual description.

In the multilingual lexical entry, the information about the syntactic and semantic behavior of an entry is constrained (adding or deleting semantic and/or syntactic information) by means of a set of transfer conditions that allow to create correspondences between language pairs. In other words, all information concurring to define a syntactic structure or a word meaning from a monolingual point of view can be exploited for multilingual requirements and, together with the transfer conditions, can be regarded as **basic notions.**

A general description of the basic notions will be provided by means of examples highlighting the role of basic notions in the multilingual perspective. They will be also described in terms of their constitutive sub-elements, thus paving the road towards a more formal definition of these objects (section 6.3).

The main input to this work comes from the previous experiences:

- The Recommendations on Subcategorization (available for browsing and download at http://www.ilc.pi.cnr.it/EAGLES96/synlex/synlex.html) and on Lexical Semantics (http://www.ilc.pi.cnr.it/EAGLES96/EAGLESLE.PDF) proposed by EAGLES, where already emerged a very large set of agreed-on information
- The syntactic and semantic layers of the PAROLE and SIMPLE lexicons, which (built-up with the flexible and harmonized GENELEX model, uniform criteria and types of information for twelve EU languages), can be seen as plurilingual lexicons, allowing cross-language linking.
- The ISLE Survey of main approaches towards bilingual and monolingual lexicons (Calzolari et al., 2001a), which provides an examination of linguistic phenomena crucial to sense distinction and to the selection of the correct translation equivalent.
- The work on Sense Indicators (cf. 5.2).
- The experience gained from the creation of mono and bilingual lexical entries (cf. 5.1).

6.2.3 Basic Notions for Syntax

Dealing with multilingual phenomena implies the treatment of the numerous linguistic facets concurring to determine the behavior of an entry in its syntactic context. A special role in this regard is played by the notion of syntactic frame, the structure that contains the syntactic arguments of an entry, their phrasal realization, the entry itself, and its probability to appear in a corpus with a specific syntactic context. The notion of syntactic frame will be introduced here, together with its sub-elements and attributes.

6.2.3.1 SYNTACTIC FRAME

The possibility to express in an explicit way the information inherent to the complementation pattern of a lexical entry is crucial for the implications in a multilingual framework.
The notion of subcategorization has been the object of investigation of a previous phase of EAGLES and its results constitute the EAGLES Recommendations on Subcategorization of which the PAROLE lexicon architecture[13] is an example of instantiation (http://www.ub.es/gilcub/SIMPLE/reports/parole/parole_syn/parosyn.html).
It corresponds synthetically to a set of possible syntactic structures (the head and its syntactic arguments) associated with an entry (typically a verb, but also a so-called predicative noun, an adjective or an adverb). Information about subcategorization can be expressed by means of a list of sub-elements and in this sense can be considered as a complex basic notion. Sub-elements are:

1. A list of slots/positions representing the syntactic arguments (mandatory or optional) and their phrasal realization;

[13] The terminology comes from EAGLES. In the PAROLE-SIMPLE specifications the notion is termed Description.

2. Categorial and morphosyntactic constraints concerning the lexical unit being described (the *Self* in EAGLES terminology);
3. Surface order information;
4. Frame probability.

Under a general perspective, "Subcategorization is concerned with the lexical specification of a predicate' s local phrasal context" (Sanfilippo, 1996). The two notions of subcategorization and argument structure are strongly interconnected: they both are at the heart of the correspondence between syntax and semantics. They have a strong discriminating power in the translation selection, giving rise to different translation equivalents on the basis of the different thematic roles and semantic characterization a syntactic position can take.
The notion will be presented here only at the level of syntax, focussing on how the complementation pattern of an entry can be used to address the translation in the right direction. The correspondence between syntax and semantics will be dealt with later, after the introduction of the basic notions for semantics.

Different syntactic readings of the same lexical unit may have implications from a multilingual point of view. In this sense, the absence of a complementation pattern should also be considered a kind of syntactic description by itself which may have a discriminant power vs. another frame-bearing reading of the same lexical units. Let us consider the typical polysemy "abstract vs. concrete" nouns incur into: the 0-frame noun, preferably, bears a concrete reading, whereas the frame-bearing noun goes towards an abstract sense. From a multilingual perspective, the different constructions may also imply different translations. For example, the Italian *velo* gets different translations according to the different complementation patterns (0-frame *vs.* frame-bearing construction):

un abito di velo (*a voile dress*) *vs.* un velo di tristezza (*a veil of sadness*)

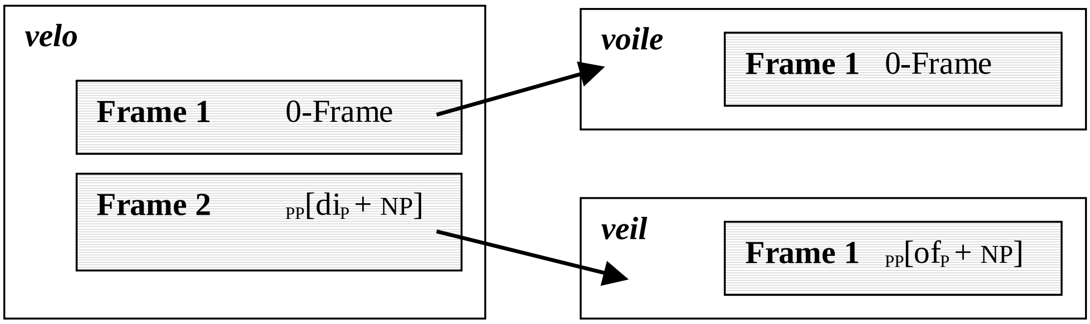


Figure 14

bassa marea (*low tide*) *vs.* una marea di gente (*a stream of people*)

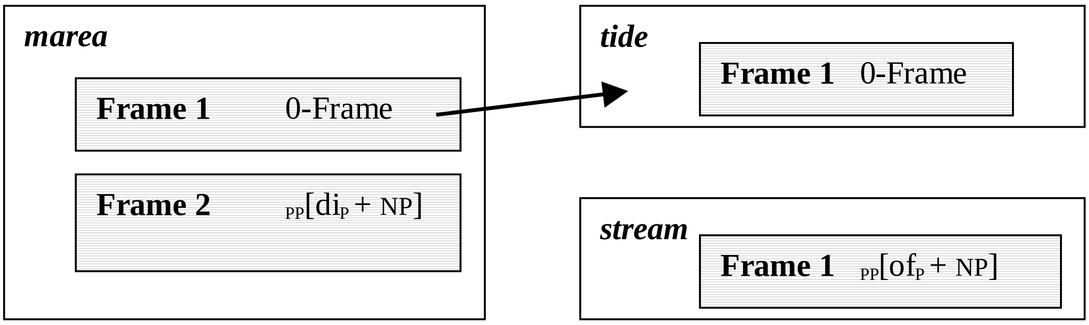


Figure 15

6.2.3.2 REGULAR SYNTACTIC ALTERNATIONS AND FRAMESET

The FrameSet has been proposed in the Report on Subcategorization by EAGLES among the set of recommended information, with the aim of explicitly relating together different surface regular alternations associated with the same deep structure (or predicate). At representational level, the mechanism of FrameSet allows to collect together, in a same syntactic entry, systematic alternations of frames that do not imply differences in meaning, by relating the "underlying structure" with the "surface structure", and specifying the rules that link the slots or slot fillers of the alternating structures. Phenomena generally dealt with by the FrameSet are:

- causative/inchoative alternations
- locative alternations
- different structures of symmetric verbs
- intransitive/transitive *vs.* reciprocal alternations

The figure below shows how the device works for e.g. the causative/inchoative alternation. A special object allows to relate the two slots that correspond each other in the two alternating frames: in *RelatedElement1*, the slot1 of the causative structure is declared as the element to be put in relation, whereas *RelatedElement2* contains the second term of the relation, i.e. the slot0 of the inchoative reading. In other words, it is simply declared that the object of the transitive reading corresponds to subject of the intransitive one.

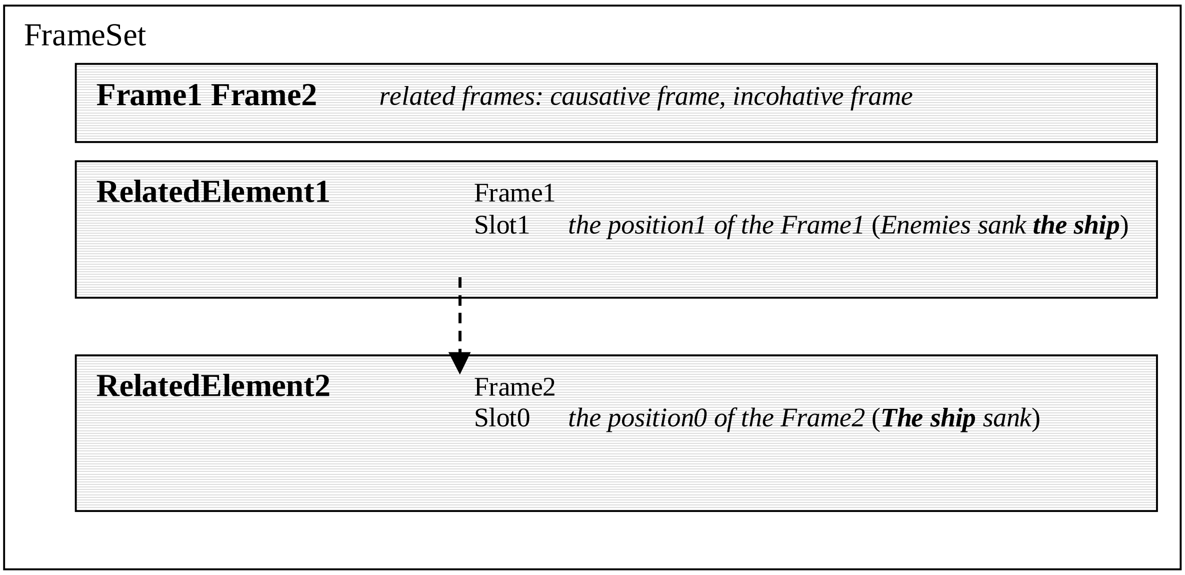

FrameSet
Frame1 Frame2
related frames: causative frame, incohative frame
RelatedElement1
Frame1
Slot1
the position1 of the Frame1 (Enemies sank the ship)
RelatedElement2
Frame2
Slot0
the position0 of the Frame2 (The ship sank)

Figure 16

6.2.3.3 SLOTS

Slots are the subcategorized elements of the syntactic frame (the *syntactic positions* in the GENELEX/PAROLE terminology) and can be described in terms of:

- **Categorial and morphosyntactic information** expressing the syntactic property of a slot realization. The slot can be filled by a *terminal or non-terminal syntagma*.
  This is the place where the phrasal realization of the syntactic argument can be specified (saying for example that the first slot, Slot*0* – or in PAROLE terminology, Position*0* – is instantiated by a Noun-Phrase. etc.).
  A list of **non-terminal categories** is given in the EAGLES Recommendations on Subcategorization (Sanfilippo 1996, pp. 64-65):

S-Sentence
VP-Verb Phrase
NP-Noun Phrase
PP- Preposition Phrase
AP- Adjective Phrase
ADVP- Adverbial Phrase
DETP –Determiner Phrase
XP- Underspecified Phrase

Different surface realizations of the same position can have a strong multilingual valency: the following example shows the Italian verb *sapere* (*to know something*) that gets different English translations depending on the phrasal realization of its complements[14]:

sapere
Frame 1: Gianni sa la verità (*Gianni knows the truth*)
Frame 2: Gianni sa nuotare (*Gianni can swim*)

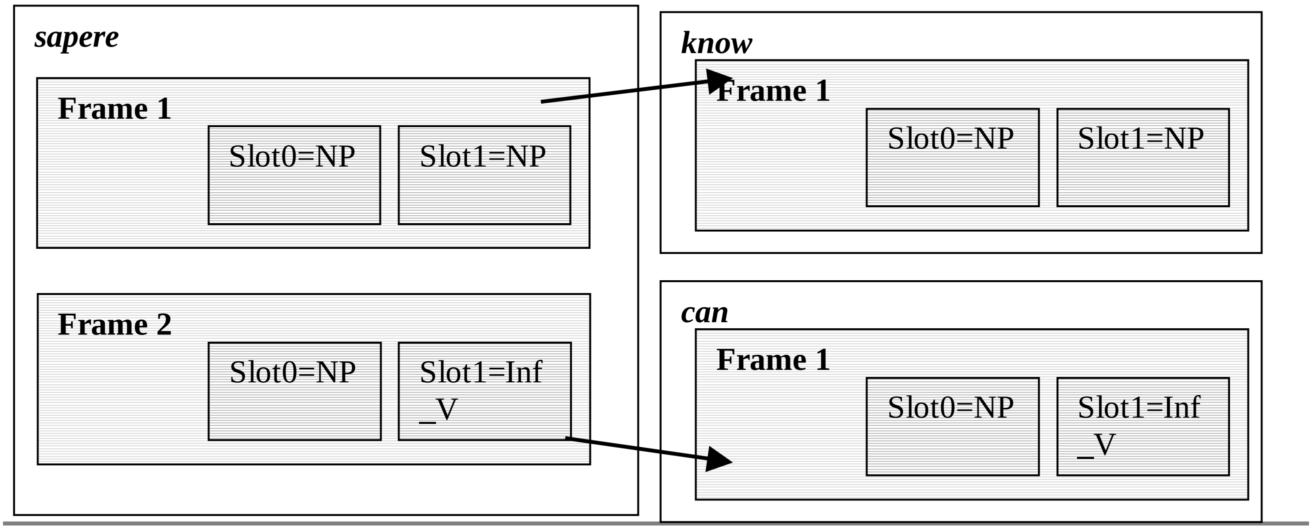


Figure 17

A slot filler can also be described in terms of **terminal categories** (the object SyntagmaT of PAROLE), for example those provided by the EAGLES Morphosyntax Group (Monachini & Calzolari, 1996):

N- Noun
A- Adjective
P- Pronoun
V- Verb
ADV- Adverb
CNJ- Conjunction
ADP- Adposition

[14] We refer here to the examples already used in the Survey of Available Lexicons (Calzolari et al., 2001a).

DET- Determine
ART- Article
INTJ- Interjection

Besides grammatical category and functions, slots can also be characterized using **restricting features**, i.e. labels that allow to specify further restrictions of morphological kind (i.e. tense, mood, gender etc…) or lexical kind (for example the lexical introducer of a prepositional phrase).

Since the same features can be also used to characterize the information about the head of the construction (the *Self* in the EAGLES terminology), their treatment will be presented in a separate section (cf. 6.2.4.5).

- **Grammatical Function** is the characteristic of a slot realization which expresses the syntactic relation linking the slot to the head it subcategorizes for.
  In the EAGLES work on subcategorization the recommended grammatical functions are a small set of few elements[15], comprising:

  -subject/complement and predicate (*necessary*);
  -direct and indirect object (*recommended*);
  -clausal components and second object (*useful*).

  The grammatical function characterizing one of the syntactic positions of the frame turns out to be a crucial notion under the multilingual point of view, since it can be constrained adding information at multilingual level and expressing, for example, a typical object or subject of a verb. Also within the Lexicographically Relevant Facts inventoried (section 5.2), *typical subject* or *typical object* turn out to be very frequent sense indicators. The following example shows the possible translations of the Italian verb *dondolare* (*to swing one's arms*, *to dangle one's feet*, *to rock the cradle*) according to the different typical objects:

[15] In the PAROLE specifications a larger set of syntactic functions, a sort of *edited union* with nearly 40 relations is available (see http://www.ub.es/gilcub/SIMPLE/reports/parole/parole_syn/parosyn.html).

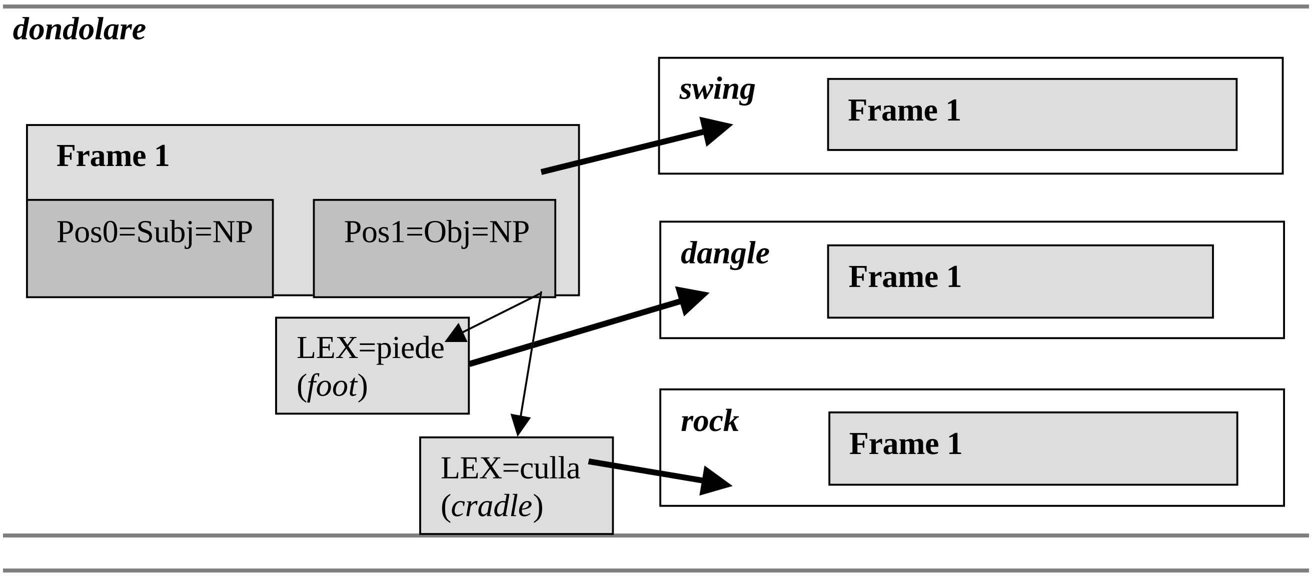


Figure 18

- **Order**

The relative order of the slots in the surface syntactic realization can be provided by a progressive number (starting from 0).

- **Frame Probability**

Frame Probability is a notion coming from the area of lexical knowledge acquisition and is not part of the previous EAGLES recommendations. As stated in Roland and Jurafsky (1998), "each lexical entry for a verb expresses a conditional probability for each potential subcategorization frame". In this sense, the lexical entry can be regarded as a vector of probabilities associated with its syntactic descriptions. Statistical information in the lexical entry is useful in a multilingual dictionary: if some subcategorization frames are more likely to occur than others, then it is possible to use this kind of information to address the translation to the most likely equivalent in the target language. The information about Frame Probability is always relative to a specific corpus and thus can be expressed by a couple constituted by an absolute number indicating the frequency of the frame (or by a percentage or an index of probability) and by the reference corpus.

- **Optionality**

In many cases, there is the need to state the optional realization of a syntactic slot within a subcategorization frame. In order to assess the optionality e.g. of a verb argument, 'nuclear' sentences should be considered , in a 'not-marked' context (since marked context can admit even the omission of traditionally obligatory complements). For the verb *to sing*, the structure *I am singing* can be considered self-explanatory, whereas, for the verb *to buy*, *you are buying* is retained as needing an obligatory direct object for the completion of the sentence[16]. Optionality, in a monolingual framework, can turn out to be a clue for sense disambiguation, e.g. a literal meaning *vs.* a figurative reading: *la legna si accese* (*incendiarsi*) *vs. Gianni si accese d'ira* (*adirarsi*). Additionally, in a multilingual perspective, this can imply different translations: *the wood caught fire vs. John blew up with rage*. The same can be true for nouns, e.g., *I lost my key*

[16] As already noted there exist some marked contexts where the verb can stand alone: let consider, e.g., *you are influenced by advertising and buy.*

(instrument) vs. to know the key (solution) to the enigma, where the abstract sense obligatorily requires the presence of the slot pp-*to.* See above for implications in translation of obligatory complements.

Restrictions on the presence/absence of slots can be also operated, the so-called *conditional optionality*:

- the absence of a slot excludes the presence of another slot : cf.
  John refuses obedience to Mary/John refuses obedience/John refuses
  but not **John refuses to Mary*
  where the absence of the direct object prohibits the presence of the indirect object.

- the absence of a slot makes obligatory the presence of another slot: cf.

  John competes with Mary for the exam/John competes for the exam/John competes with Mary
  but not **John competes*
  where the presence of one of the two slots is needed in order for the sentence to be acceptable.

6.2.3.4 INFORMATION ON THE SELF OF THE SYNTACTIC FRAME

Another sub-element of the syntactic frame is the lexical item placed in the specific syntactic environment, for which it is important to explicitly encode the part-of-speech information and all the features characterizing its morphosyntactic behavior, i.e. the auxiliary, morphological restrictions like number and gender for nouns, syntactic information like mood and tense for verbs, etc. This information is usually already described at monolingual level, but can also be added by means of a specific lexical operation in the multilingual level, when it is useful to address the translation in a specific direction. Very important is the possibility to specify complex heads in order to represent polylexical units. A complex head is something having an inner structure made of embedded positions describing the multiword components. This necessity strongly arises during the phase of entries creation, when it is important to have at disposal a device to represent in a straightforward way an entry like "make an impression" (complex head formed by *make* -verbal head- + a slot for the NP "impression").,

6.2.3.5 RESTRICTING FEATURES

The information about the syntactic frame and the syntactic behavior of an entry can be further specified by means of a set of features. In most cases, the only use of categories is not sufficient to supply the necessary information and, categories must be completed by using restricting features.

The EAGLES Documents on Subcategorization (Sanfilippo, 1996) and on Morphosyntax (Monachini & Calzolari, 1996) provide a classification of the possible types of information that can be used to refine the information already specified in the Slots and in the Self.

Features are distinguished in

- morphosyntactic
- lexical

Morphosyntactic restrictions can be imposed in the slot realization to account for

- cases that e.g. constrain a plural realization of a complement:

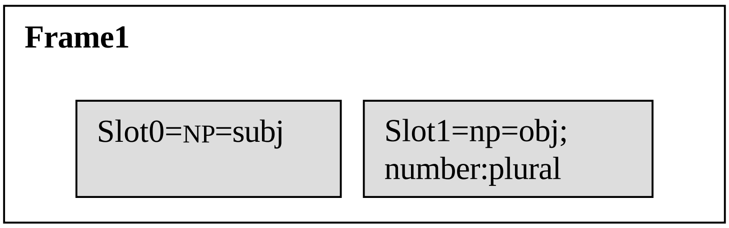


Figure 19: ex. *collezionare francobolli (to collect stamps)*

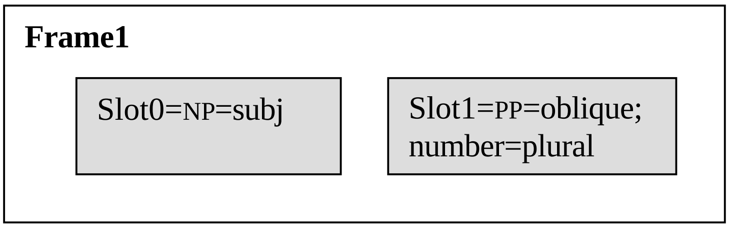


Figure 20: ex. *pullulare di stelle (to swarm with stars)*

- cases that constrain information according to the feature mood, e.g. Italian cases where the that-clause forces the subjunctive mood.

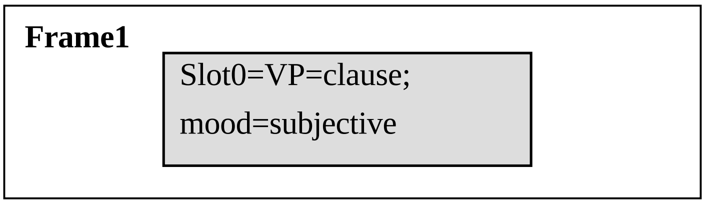


Figure 21

Beside refining information at monolingual level, this kind of information results to be crucial at multilingual level for the selection of the correct translation and also for the generation of the right context. The example below shows the mechanism of constraining the information about the number of the self in order to reach the correct correspondent in the TL (the Italian *aiuto* can be translated by *help* or *aid* depending on the number):

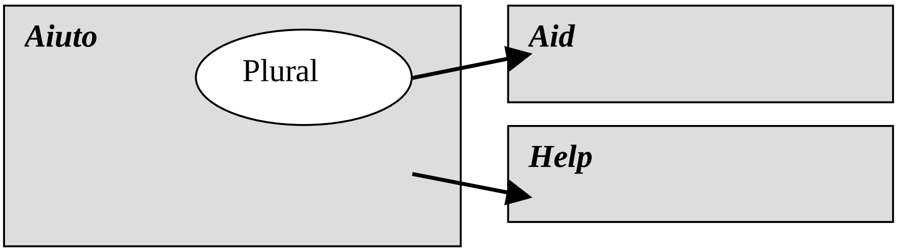


Figure 22

In the same way, the gender of the Italian *figlio* can be constrained to reach the masculine s*on* and the feminine d*aughter* of English.

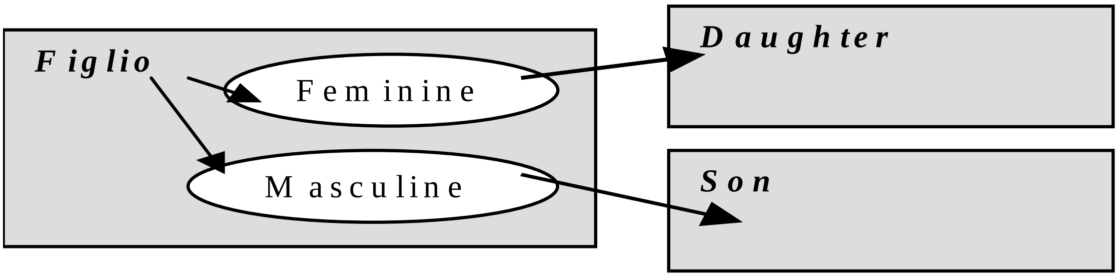


Figure 23

Lexical features, on their turn, help to describe various aspects of the lexicalization of a phrase (its preposition etc.) and are also crucial at multilingual level, since we may need to select a specific preposition within a subcategorization frame.
For example, the Italian verb *prendere* takes different English translations according to the preposition that introduces the PP: Gianni <u>prende</u> per il bavero Paolo (*John <u>seize</u> Paul by the scruff of his neck*) and Gianni prende a schiaffi Paolo (*John slaps Paul's face*).

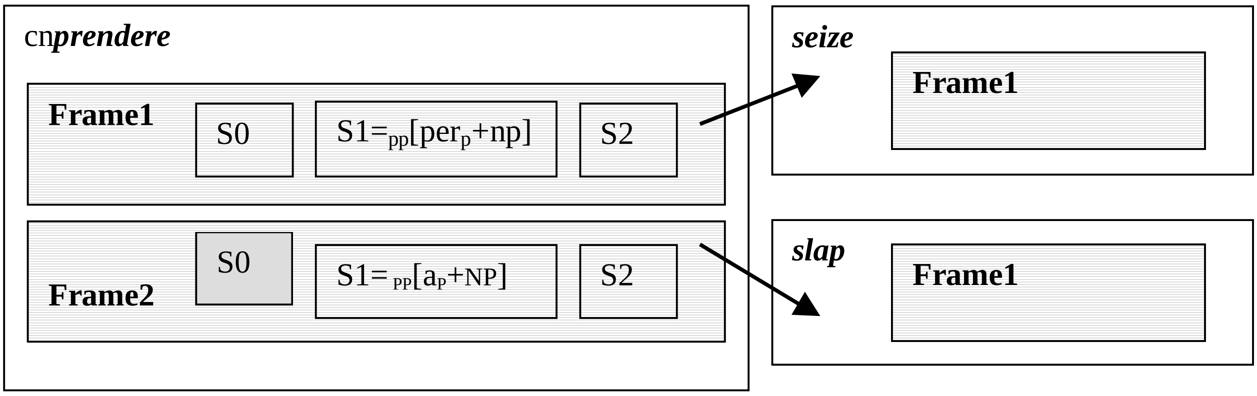


Figure 24

Control is a kind of information that can be expressed by means of features (see Sanfilippo 1996 and the PAROLE instantiation of GENELEX 1994). Control is a crucial information of a syntactic frame, since "deals with relations between two slots", e.g. an element which is understood in an infinitive clause (controlled) and a participant of the verbal frame (controller) of the governing sentence. Concretely, information can be expressed at two levels of representation, at the level of frame, a feature will specify that there is the presence of control in a syntactic frame, and special values will indicate the kind of control: subjectcontrol, objectcontrol, indirectobject control and, at the level of slot realization, where controller and controllee can be related.

Gianni$_i$ *afferma di* $\emptyset_i$ *poter venire*	SUBJCONTROL
Gianni$_i$ *promette a Maria di* $\emptyset_i$ *venire alla festa*	SUBJCONTROL
Gianni accusa Mario$_i$ *di* $\emptyset_i$ *essere un ladro*	OBJCONTROL
Gianni prega Luca$_i$ *di* $\emptyset_i$ *venire alla festa*	OBJCONTROL
Gianni chiede a Mario$_i$ *di* $\emptyset_i$ *svolgere un lavoro*	INDOBJCONTROL
Gianni impedisce a Luca$_i$ *di* $\emptyset_i$ *andarsene*	INDOBJCONTROL

In raising constructions (cf. Sanfilippo, 96, p. 81), the subject expressed in the governed sentence is “raised” as subject of the governing verb [17].
sembra che Luca sappia l’inglese (It seems Luca can speak English) ➔ *Luca sembra sapere l’inglese (?Luca seems to be able to speak English).*

Control may also have impact on sense distinction, since in some languages a difference in control switches on different meanings, cf. French *dir* and Italiano *dire* that select the sense of directive speech act (vs. declarative speech act) in presence of control on indirect object.

6.2.3.6 LINEAR ORDER

The slots of the subcategorization frame have a conventional or canonical order that can be different from the linear order of the positions in real sentences, since the surface order is not something that should be encoded in the lexicon. Anyway, as stated in the Recommendations on Subcategorization (Sanfilippo 1996.), “*for some lexical units and for some languages…some verbs may constrain the possible order of their slots or slots realizations more than others*”.
The information about linear order can be important: for example, in Spanish and in Italian, the position of the adjective as pronominal or postnominal (or both) encoded in the lexicon has consequences on the sense distinction, (i.e. *pobre hombre*/*pover uomo –unhappy, miserable man-* is different from *hombre pobre*/*uomo povero -poor, lacking money man-*).

6.2.4 Basic Notions for Semantics

At semantic level, basic information units are represented by ***word-senses***. All information concurring to discriminate senses in a monolingual framework or to direct towards a given translation in multilingual operations are regarded as basic notions. It is at the level of sense distinction that cross-language links are established and this is the reason why this level appears to be crucial in a multilingual environment.

The previous EAGLES guidelines in the area of lexical semantics have been hence re-interpreted under this perspective, trying to provide the set of information necessary to deal with multilingual phenomena. In this light, the bulk of semantic information encoded in the SIMPLE lexicons (that, built on the EAGLES recommendations, have been taken as the monolinguistic basis for the analysis carried out here) are also re-examined, integrated (with other dimensions coming e.g. from WordNet) and all wrapped up in view of the MILE. Other realities have been taken into account, since the notion of *word meaning*, which is central to semantics description, is not uncontroversial.

[17] In Italian, *subject-raising* structures only exist.

In the lexicographic tradition, the word meaning is the *sense*, the unit resulting from the subdivision of the lemma in its readings. In lexicons *à la* GENELEX (or SIMPLE), the word meaning is represented by the SemU – the Semantic Unit – corresponding to the traditional notion of word sense and constituting the nuclear building block of the whole semantic description. It is the semantic unit that is linked to a given ontological type, it is the semantic unit that the semantic frame is associated to, and it is the semantic unit that, alternatively, works as the target and the source of all semantic relations. A different modality of representation resorts to the *synset*, the *set of synonyms* that constitutes the building block in WordNet (Fellbaum, 1998) and WordNet-like kind of resources (Vossen, 1999). During the years, WordNet has become an outstanding reality for the lexicon community, with WordNets dedicated to dozens of languages and used in a wide variety of applications. Thus, it is important to take WordNet and its basic structure into consideration, ensuring that all the already encoded resources could be easily mapped into the final ISLE recommendations.

In the same way as for the syntactic side, in semantics, basic notions can be of two types: simple or complex. A simple notion is simply constituted by the notion itself (e.g. Domain), whereas the complex one subsumes and can be described in terms of other sub-elements (e.g. the semantic frame subsuming other elements, such as Predicate, Arguments, Roles, .., each of them working as basic notion).

6.2.4.1 SEMANTIC FRAME

This is a complex notion, that specifies the predicative argument structure of a lexical unit described in terms of the following types of sub-elements: the predicate, which on its turn is described by means of a list of arguments, their semantic role and the selectional restrictions the predicate operates on them. This notion "incorporates most of the lexical semantics elements, since predicates are often the 'kernel' of propositions" (*Sanfilippo*, 1999). It is an important element in monolingual perspective, in lexical resources and for applications, thus playing a crucial role in establishing links between the syntactic and the semantic levels. In SIMPLE, where the semantic frame is recommended and instantiated with a very high degree of detail (cf. Lenci *et al.* 1999, p. 46), information about the type of link between the predicate and the unit at hand is also provided. This information can have repercussions on cross-language linking.
In a multilingual perspective, the semantic frame is the place where many operations necessary to go from L1 to L2 occur: all information connected to it helps such operations.

6.2.4.1.1 Predicate

The information about the predicate is relevant mainly for verbs, predicative nouns, adjectives, prepositions and adverbs. The approach to predicates can be of two types: multilingual, as language-independent primitive predicates, or monolingual, as language-dependent lexicalized predicate. On the one hand, 'abstract' predicates to be shared by homogeneous classes of semantic units across languages could acquire a kind of "interlingua" valency (PROPERTY_OF, which could be linked to all nouns indicating property, such as *bellezza, beauty, beauté; altezza, height, hauteur,* …independently of lexicalization in every language). EAGLES recommends and SIMPLE instantiates language-dependent lexicalized predicates which present "the advantage of reducing the complexity of the linking with syntax" (Lenci et al. 1999, p.46). Predicative entries are ascribed a *semantic predicate,* being provided with the so-called predicative representation. The approach followed in SIMPLE for the selection of predicates foresees that members of a whole derivational paradigm are all linked to the same predicate. It follows that different semantic units may share the same predicate in the predicative representation: e.g. the verb *destroy* and the nouns *destruction* and *destroyer* all point to the PredDESTROY; similarly, the verb *employ,* and the nouns *employment, employer* and *employee* are linked to the PredEMPLOY.

The *type-of-link* is the place where the different relations holding between the semantic unit and the assigned predicate are reflected:

- Verbal lexical units *employ* and *destroy* present with respect to their predicate a ‘MASTER’ type-of-link, which stands for ‘the priviledged lexicalization of the predicate’;
- *employment* and *destruction*, on their turn, constitute EVENT NOMINALIZATIONS (whose surface realizations instantiate all the arguments of the relevant predicate).

The fact of having the verbal and deverbal noun structures linked together via the predicative representation could be of extreme utility in order for, e.g., the two different surface realizations of the same predicate be recovered in translations from a language to another[18].

- *Employer* and *employee* are, respectively, AGENT and PATIENT NOMINALIZATIONS of PredEMPLOY. Within the type of link there is also the possibility to specify that in both nominalizations the phenomenon of ‘argument absorbtion’ takes place, i.e. *employer* absorbs in the lexical head the ARG0:agent, whereas *employee* encapsulates ARG1:patient.
- INSTRUMENT NOMINALIZATIONS and locatives (OTHER NOMINALIZATIONS) are ascribed the relevant predicate as well, cf. *mixer* that incorporates ARG2:instrument of the PredMIX and *breeding* that realizes ARG2:location of the PredBREED.

6.2.4.1.2 Arguments

The notion of predicate involves the specification of the number and type of arguments. Arguments as well as predicates are ‘lexically driven’, so each predicate has its ‘own’ arguments. Determining the list of arguments involved in a predicate is not a trivial task. As an example, SIMPLE states that the choice of the number of arguments for a predicate has to be determined on purely semantic grounds: it is perfectly possible for a semantic argument not to be mappable to any syntactic position, and, conversely, it is perfectly possible for a syntactic position to remain unlinked to any argument.

At multilingual level, arguments represent a critical notion, since most of the operations to go from L1 to L2 seem, principally, to affect aspects of the syntactic facet connected to a semantic frame, the number of arguments involved in Frame1 and Frame2, the order of the slots filled at the level of surface syntactic realization.

6.2.4.1.3 Thematic Roles

They specify the semantic links between the head (predicate) and the grammatical functions it governs (arguments) and it is on the basis of the recognized roles that the argument structure can be defined. E.g. the semantic frames of “giving”, “putting” and “cutting” can be recognized as trivalent structures:

donare (to give) ARG0-Agent ARG1-Patient ARG2-Beneficiary
mettere (to put) ARG0-Agent ARG1-Patient ARG2-Locative
tagliare (to cut) ARG0-Agent ARG1-Patient ARG2-Instrumental

EAGLES guidelines on lexical semantics provide a set of very basic (commonly used) thematic roles:

- Agent
- Patient
- Experiencer
- Location

[18] The event linked to the PredDESTROY can be, indeed, instantiated both as *la distruzione della città da parte dei nemici* (the destruction of the city by the enemies) and *i nemici distruggevano la città* (enemies destroyed the city).

- Instrument

They are crucial in cross-lingual operations, since roles can be assigned different surface realizations and positions in frames depending on the syntactic peculiarities of different languages, but, remaining unchanged in deep realizations, can act as a clue to generate the correct translation equivalent.

Pred*give*: ARG0-Agent ARG1-Patient ARG2-Beneficiary
Gianni dà un libro a Maria
John gives Mary a book

Cf. the multilingual section (6.2.6) for exemplifications.

6.2.4.1.4 Selectional Restrictions

Selectional restrictions should rather be intended as *selectional preferences* (cf. Sanfilippo *et al.,* 1999, Lenci *et al*. 2000 and Calzolari *et al*., 2001a) i.e. as arguments which are preferably selected by a predicate. Selectional restrictions on an argument can be specified in terms of the following types of information:

- *Semantic Type*, taken from the list of semantic types that form the Ontology (see the section on Semantic Type);
- *Features* or *Notions*, e.g. a set of semantic types (Human Animal, i.e. the ∪ of the set of Humans and the set of Animals), a semantic type plus feature(s) (Human +FEMALE) .
- *Semantic Unit*: for instance, *bark* has a two-argument semantic frame, where the second is restricted to *dog* (where *dog* should include all instances of class DOG).
- *Synsets*: restrictions can be enforced also by means of a group of admitted synonyms[19].
- *Collocations:* restrictions can involve a lemma typically accompanying the unit at hand.

Restricting the predicate's argument by means of semantic features allows to overcome cases in which the use of other expressive means, e.g. semantic types, seem to fail in capturing the full range of arguments, being, alternatively, too wide or too restrictive[20]. Features, which cut across the type hierarchy, allow in fact to capture a more suited set of lexical units and are considered more powerful in identifying preferences: cf. the restriction on patient of the Pred*eat*, that excludes vegetals and fruit if expressed with the type Food, whereas captures also other semantic units distributed over different semantic types (Vegetal, Fruit, Vegetal_entity, Substance, Natural_Substance …) if expressed by the feature [+edible] (see below for the use of distinctive features).

6.2.4.2 SYNSET

The *synset* is the set of synonyms that plays the central role of *lexical concept* in WordNet. Following psycholinguistic assumptions, the idea is that the human lexical memory is organized around concepts that words can be used to express. The same meaning can thus be carried by more than one word and represented by the group of those words themselves.
This is an important shift from the lexical organization introduced so far: the synset can be viewed as a set of *senses* of different lemmas (the *variants*, in the EuroWordNet terminology, the SemUs in

[19] Even if it should be taken into account that not always members of a same set of synonyms can be perfectly interchangeable.

[20] Selecting the type Human for the agent of the Pred*eat* excludes Animal, whereas Living_Entity covers also undesiderable Vegetal_Entity.

Genelex-Simple terminology) grouped on the basis of their reciprocal synonymy. The following list of word senses are examples of two actual WordNet1.6 synsets obtained with the search word *home*:

{*dwelling, home, domicile, habitation*} - a physical structure that someone is living in
{*family, household, house, home, menage*} - a social unit living together

The synset is the node of the semantic net, since it works as an anchor for every semantic relation (for a description of WordNet and WordNet-like resources cf. 4.2.2)

The whole wordnet-like architecture can be represented on the basis of the following elements:

- The synset with one or more synonyms (variants, senses, SemUs.) as sub-elements and characterized by the following attributes:
    - POS indicator (*mandatory*)
    - Gloss (*optional*)
    - Example (*optional*)
- a list of one or more relations. The relations can be of different types, representable by means of different attributes: monolingual semantic relations, equivalence crosslingual relations and plug-in relations[21] between generic and domain-specific wordnets.
- Features providing the semantic and ontological types.

6.2.4.3 FEATURES

6.2.4.3.1 Semantic Type

Semantic type appears to be a crucial notion, since it establishes a link between a word-sense and an ontological type system which is used to classify senses themselves, thus allowing to assign it to a specific position in the nodes of the type hierarchy: *dog* [Animal ← LivingEntity ← ConcreteEntity ←…]. In cases where senses are not defined on the basis of an ontology, the semantic type can be also obtained via semantic hyperonymic relations with another word-sense, *dog* isa *animal*.
Consensus over the importance of this notion is easy to find: the semantic type of a word sense is a mean to discriminate among other possible senses of the same lemma. Looking at well-established practices in computational lexicons or Machine Readable Dictionaries, all of them make use of it (Calzolari et al., 2001a). This notion is considered as *required* by SIMPLE (Lenci et al. 1999, p.37), i.e. it is part of the core information included in the minimal requirements for computational lexicons at semantic level (For a complete list of the SIMPLE Semantic Types, cf. Appendix D).

In a multilingual perspective, the usefulness of the semantic type of a word in sense distinction (for analysis and generation) is uncontroversial.

L'incidente [Event] *colpì Maria*[Human] ➔ *to impress*
Il tifone[Phenomenon] *ha colpito il Giappone*[Location] ➔ *to damage*
L'aggressore [Human] *colpì la vittima* [Human] *col pugnale* [Instrument] ➔ *to hit*

6.2.4.3.2 Domain

Information about domain is available in most dictionaries and lexicons. It results to be a critical notion, since it has a discriminant power in sense distinction and can impose semantic constraints in translation selection. Cf. e.g. the different translations in Italian of Eng. *mouse,* resulting from different domains It. *topo* and It. *mouse*.

[21] As instantiated in the ItalWordNet databases (Roventini et al, 2002).

For the importance of domain information to retrieve a whole semantic field (e.g. that of Cuisine), see Ruimy *et al.* 2002.

6.2.4.3.3 Distinctive Features

The use of distinctive features can allow to refine the semantic information, thus enriching the information provided by means of the semantic typing of an unit. Such features, indeed, which cut across the type hierarchy allow to capture meaning dimensions which are orthogonal to the ontology and are not expressible resorting only on it. This is the case of e.g. *edible entities* which are not part of the node Food, but belong to other ontological nodes, (e.g. Natural)Substances, Vegetable and Fruit (these two last subnodes of Living_Entities, etc.) and do nor inherit the characteristic of being edible. The use of the feature [+edible] allows to restore this information, which is useful, in monolingual perspective, for retrieving all edible entities sparsed over different semantic type, in view of the enforcement of correct selectional restrictions (see above). In cross-lingual operations, the use of distinctive features acquires discriminating power, allowing to account for the different translations of e.g. the Fr. *avocat* into Eng. [+edible] *avocado* vs. the [+human] *lawyer*.

6.2.4.4 SEMANTIC RELATIONS

Together with the above expressive devices, the semantic purport of an entry is also represented by means of semantic relations between two semantic units (senses). Information that traditionallly is committed to relations consists in meronymy – *part_of* (finger, hand) –, and its inverse relation holonymy – *has_part* (carburettor, car) –, antonymy, with its variuos types of opposite relations – (true, false); (hot, cold) – as discussed in Cruse, 1986. The utility of such dimensions in various types of applications is carefully reported in the EAGLES Recommendations on Lexical semantics (cf. Sanfilippo, 1999, p. 238).

In the framework of the SIMPLE experience, traditional Qualia roles of the Generative Lexicon (Pustejovsky, 1995) have been implemented as relations between SemUs This allowed lexicographers to represent the richness of semantic relations in natural language and, at the same time, to capture the essence of a word meaning. In addition, the set of Qualia roles has been made richer and simultaneously stricter. Richer because each of the four Qualia roles has been represented in the form a relation, which is in turn the top of a hierarchy of other more specific relations. Stricter in that the enlarged set of relations allow to capture more fine-grained relations holding between different senses. These hierarchies of relations (specifically 64 semantic relations) within the four Qualia have been called *Extended Qualia Structure*, (cf. Lenci et al. 1999, pp. 59-71,). Qualia relations, combined together, characterize, indeed, semantic types of different degrees of complexity and concur to maintain the (Qualia) structure of a semantic type. Relations have been also given a weight, depending on their being type-defining with respect to a semantic type or not.
Derivational relations (*beauty*; *beautiful*) and regular polysemous classes (animal, food: *lamb*$_1$, *lamb*$_2$; substance, color: *porpora*$_1$, *porpora*$_2$ - Eng. *purple*, *crimson* -, etc.) have been expressed by means of relations between semantic units as well.
It is worth noting that the SIMPLE relation system as it is conceived is very flexible and extensible suitable to hold other kinds of relations coming from other semantic approaches, such as e.g. Frame Semantics or Mel'cukian lexical functions.

So far, we have talked about relations between units of sense but semantic relations can also be established between synsets, as in the WordNet model.

In general, we can talk about "relational models of semantic representation" or "relational dimension of semantic representation". In relational models relations can hold between word senses (or Semantic Units) or set of synonyms (SynSets).

6.2.4.5 COLLOCATIONS

Collocations, which EAGLES defines a kind of "word co-occurrence relations" (cf. Sanfilippo *et al.* 1999, p. 240), are crucial to define the semantic purport of a lexical entry which selects a particular meaning when it co-occurs with a given word. In collocations, the way words go together seems idiosyncratic and unpredictable: the selection operates at the lexical level rather than at general semantic level. This has a particular impact in multilingual operations in order to arrive at the correct translation equivalence in another language. Collocations can by their nature be encoded by means of the expressive device of relations, where the typical collocate of a word is the target of the relation[22]. EAGLES provides a set of information generally necessary to be specified for collocations (cf. Sanfilippo *et al.* 1999, p. 245): direction, word-distance, dependency, dependency type, probability.

6.2.5 Linking Syntax and Semantics

The type of notion dealt with in this section refers to one of the most crucial aspects of computational lexicons, which goes by the name of *linkage of syntactic and semantic levels*.
The PAROLE/SIMPLE model tackles this task, offering a very effective solution. Once predicative entries are ascribed a semantic predicate, the operation of linking between syntax and semantics is made by means of a battery of *mapping rules* that correlate the *semantic frame* pointed by a semantic unit to the *syntactic frame* the latter is associated with, indicating how *semantic arguments* and *syntactic slots* correspond each other, i.e. how arguments are instantiated in the surface.
Rules to map the semantic predicate onto its possible syntactic surface instantiation(s) have to cope with the following cases of correspondence:

- relations of isomorphism, where slots and arguments correspond to each other in number and range (mono- bi-, tri-, tetra- valent ISOMORPHIC correspondences: ARG0-SLOT0; ARG1-SLOT1 …),
- relations between slots and arguments appearing in crossed order (CROSSED correspondence: cf. *destroy* and *destruction*: ARG1-SLOT0; ARG0-SLOT1)

To give but an example of the usefulness of the mapping rules and just a flavour of how they work, a case of regular dative alternations is taken into consideration: starting from the same semantic predicate, the two alternating surface realizations can be reconstructed by way of the appropriate mapping rules.

1. *John gave a book to Mary.*
2. *John gave Mary a book.*

The two syntactic frames are associated with two different syntactic units:

- syntactic-unit$_1$ corresponds to an NP NP PP-*to* syntactic frame, cf. in (1)
- syntactic-unit$_2$ corresponds to the NP NP NP variant, in (2).

[22] The SIMPLE model allows to encode collocations as relations between semantic units: *collocates* (*potente*, *farmaco*) means that the typically accompanying noun of the adjective *potente* is *farmaco*, where *potente* = *effective* and *farmaco* = *drug*.

Both are associated to the same semantic unit <give>[ChangePossession] which points to the predicate PredGIVE(Arg0:agent, Arg1:patient, Arg2:beneficiary). This semantic frame results implicitly linked, via the same semantic unit, with the two alternating syntactic frames, as shown below.

Syntactic frame	Semantic Unit	Semantic Frame
give1:Slot0:NP,Slot1:NP Slot2:PP-*to*	give[ChangePossession]	PredGIVE(Arg0:*agent*,Arg1:*patient*,Arg2:*beneficiary*)
give2:Slot0:NP,Slot1:NP Slot2:NP	give[ChangePossession]	PredGIVE(Arg0:*agent*,Arg1:*patient*,Arg2:*beneficiary*)

Table 22

Different mapping rules will account for the differences in correspondence between the predicative structure and the two possible surface instantiations: the PredGIVE, on the one hand, is associated with syntactic-unit$_1$ by an *isomorphic* correspondence, where Arg0 is linked to Slot0, Arg1 to Slot1 and Arg2 to Slot2, on the other hand, will be linked to syntactic-unit$_2$ by a *crossed* correspondence, in which Arg0 is linked to Slot0, Arg1 to Slot2 and Arg2 to Slot1 (cf. GENELEX 1994).

More problematic cases to deal with correspondences can be:

- syntactic slots that do not map onto predicate arguments, the case e.g. of adjuncts which are part of the syntactic frame but extraneous to the semantic one (REDUCED correspondence) or, conversely,
- semantic arguments that do not appear in surface realizations (e.g. 'Meteorological' predicates [*snow*] *snowed*) or can be lexically encapsulated (AUGMENTED correspondence).

In multilingual perspective, argument encapsulation has interesting implications, when dealing with cases of predicates which behave differently[23], across languages wrt this phenomenon, cf. Eng. *to funnel* – It. *versare con l'imbuto* and Eng. *to hammer* – Fr. *enfoncer avec un marteau*. The translation equivalences between English and Italian and English and French, respectively, can be obtained following two different strategies: the first – in an *interlingua*-like approach[24] – is to make use of an abstract and language-independent predicative representation – where the arguments contained in the lexical heads *to funnel* and *to hammer* appear esplicited – as a bridge between the two instantiations, exploting the set of mapping rules to reconstruct the correct surface realizations: Pred*funnel* (ARG0agent ARG1patient/+liquid ARG2instrument/funnel), and Pred*hammer* (ARG0agent ARG1patient ARG2location ARG3instrument/hammer). Another possibility – more typical of *transfer*-based models – is to deal with the same phenomenon in a more concrete and practical way, putting in correlations the two monolingual syntactic and semantic frames in L1 and L2 and specifying the set of tests and actions necessary to go from L1 to L2, e.g. ADDARGUMENT below.

[23] Some of them present, indeed, diverging surface realizations, being the argument implicit or contained in the lexical head in one language – 'shadow' argument – and explicit in another.

[24] *Interlingua*-based approaches tend to deal with translation at a high level of abstraction, decomposing meaning in more and more simple elements, in such a way that it is as independent as possible from language instantiations. The level of predicate is deemed to constitute the suitable place for such decomposed representation of meaning and the SIMPLE correspondences can play as rules for generating, from abstract conceptual/semantic descriptions, the correct concrete realizations in the different languages.

6.2.6 Basic Notions for the Multilingual Layer

The aim of ISLE is to provide a common model to represent multilingual content. We will introduce the current main approaches to multilinguality, in order to circumscribe the linguistic and representational issues we have to deal with. A step towards the definition of such a common model is the identification of the basic notions for the multilingual level. If at the monolingual level basic notions mostly concern "static" lexical objects (such as syntactic slots, semantic arguments, restricting features etc.), from a multilingual perspective basic notions involve the set of operations that uses these very lexical objects as *arguments*.

In what follows, a brief overview of the main approaches to the representation of multilingual content is given, together with the set of operations that can be considered the basic notions for the ISLE multilingual layer.

6.2.6.1 APPROACHES TO THE REPRESENTATION OF MULTILINGUAL CONTENT

In Dorr *et al*. (1999) three different MT approaches are surveyed: i) direct, ii) transfer, iii) interlingual.

The **direct architecture** is based on the simple *word-to-word* replacement, an approach that has scarce results from the point of view of syntactic correctness and word sense disambiguation, but can be of use to perform translations in a **terminological** framework.

The **transfer approach** exploits the syntactic and semantic representation of the source and the target languages, using a set of trasformational operations that allow to go from L1 to L2. The goal is to preserve the correct syntactic context and to resolve many cases of syntactic and semantic ambiguity.

The **interlingual approach** is based on the idea that translations from SL to TL should pass through a language independent representation. This approach requires a deep semantic analysis in order to specify the interlingua and does not need transfer rules since the representation is independent of the source and target languages[25].

In Dorr *et al*. (1999) the three different approaches are represented by means of a "pyramid diagram", where the types of MT systems are described moving from the pyramid base, represented by the *direct transfer*, to the vertex, the *interlingual* approach, passing through the syntactic and semantic types of *transfer*. The results of these three approaches are different, but the "right approach" has to be evaluated taking into account many factors: i) the type of MT applications, ii) type of text to be translated (text representative of a domain-specific/terminological lexicon or of a general lexicon), iii) the possibility to make post-editing work on the result, iv) time/expence constraints, v) number of language pairs to create.

6.2.6.2 ISLE APPROACH TO MULTILINGUALITY

If the absolute "right approach" to the multilinguality issue does not exist, the ISLE task is to provide a way to represent the lexical information used within these three different models in order to give the user the possibility to implement the preferred one.

The most important reference work for our analysis is the "Rapport sur le MULTILINGUISME" of the GENELEX Consortium (1994). In GENELEX, multilinguality is dealt with as the *natural evolution* of the monolingual model that has come to maturity. The GENELEX approach is

[25] Anyway, mapping rules are needed to generate from the interlingual representation the correct surface realizations.

basically *transfer-based,* i.e follows a model where a transformation of the source language representation into a suitable target-language representation is performed.

The framework of OLIF (Thurmair, 2000), the interchange format used in many industrial MT systems, has been taken into consideration as well and its set of transfer operations will be part of the operations described in the ISLE multilingual layer. It is worth reminding here that the set of official OLIF data categories - on which transfer operations (*restrictions* and *structural changes*) work - is fully compatible with the information recommended and instantiated by SIMPLE.

ISLE tries to extend the GENELEX model towards the definition a more flexible framework where different approaches can be instantiated, in particular opening the door to an interlingual approach. With respect to the objects presented in the GENELEX multilingual layer, "new" basic notions have been introduced coming from the monolingual layers, to be exploited at the multilingual level as well, i.e. the *synset* – that can be used in cross-language correspondences – and the semantic relations – on which the transfer mechanism operates in the same way as on other notions.

Even if ISLE takes inspiration mostly from a transfer-based multilingual model, in the model proposed it should be possible to represent and instantiate also a more elementary and a more conceptual/abstract multilingual model:

- the direct transfer architecture can be instantiated recurring to the simplest and immediate correspondence, i.e. that between morphological units;
- the interlingual approach to translation can be implemented, exploiting and specializing the semantic/conceptual level: the monolingual notion of lexical predicate can be extended to a more abstract notion of non-lexicalized predicate (see below, Fig. 29).

ISLE approach to multilinguality, however, is basically based on transfer and bilingual correspondences: the monolingual lexicons can be viewed as repositories that work as the *pivot* on which the single bilingual modules are based. It is in the multilingual layer that the lexical correspondences are established, resorting to the monolingual descriptions, linking together pairs of semantic lexical units, syntactic structures and semantic frames of monolingual entries. All the linguistic basic notions introduced in the previous sections (6.2.3, 6.2.4) can be the objects which the transfer rules work with, providing an easy way to implement the transfer architecture.

At *multilingual level* two sets of notions can be identified:

- multilingual correpondences
- operations that can be used in the test and action mechanism.

6.2.6.3 MULTILINGUAL CORRESPONDENCES

The first set of notions includes the multilingual correspondences, that intervene in the linking process of monolingual lexical objects. Correspondences should be possible between:

- morphological units pairs.

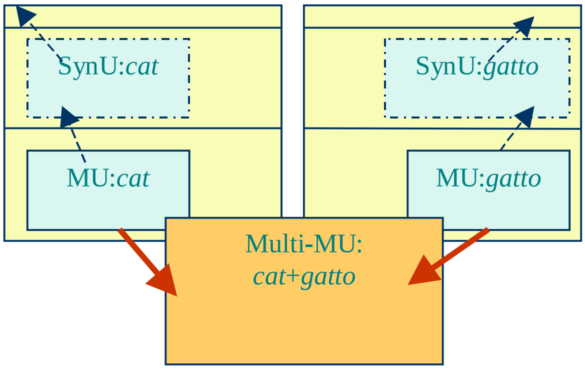


Figure 25 : Simple Direct-like correspondence using Morphological Units

- syntactic unit pairs: it allows to relate two syntactic units independently of their semantic realization. Sub-element of this kind of correspondence is the correspondence between each slot of the SL and TL syntactic frames.
  - slot pairs: it specifies the correspondences between slots of the descriptions linked to each syntactic units. It should be possible to constrain or prohibit the realization of a slot, to force it to a given syntagma. The syntagma, on its turn, should be constrained and new slots added to the already existing list of slots and again constrained.
- semantic unit pairs: when a correspondence is established between SL and TL semantic units, all the syntactic units connected to them are related, and implicitly, via the correspondence between syntax and semantics, their syntactic frames are linked as well. When predicative semantic units are put into correspondence, obviously their respective semantic frames are related as well.
- predicate pairs: this correspondence allow to associate the predicates of each language, independently of the semantic unit(s) they are pointed by and, hence, independently of the semantic frames they are linked to.
  - argument pairs: it specifies the correspondences between arguments of the semantic frames of the SL and TL. It should be possible to add a semantic feature in order to better specify the argument or operate a constraint in order to cover the semantic gap, if any, between two elements in correspondence. It should be possible also to specify optional arguments which do not present any correspondence in the other language, or, conversely, to add arguments.
- mixed pairs of semantic and syntactic units: allows to exactly specify which syntactic descriptions are linked for a given lexical meaning.

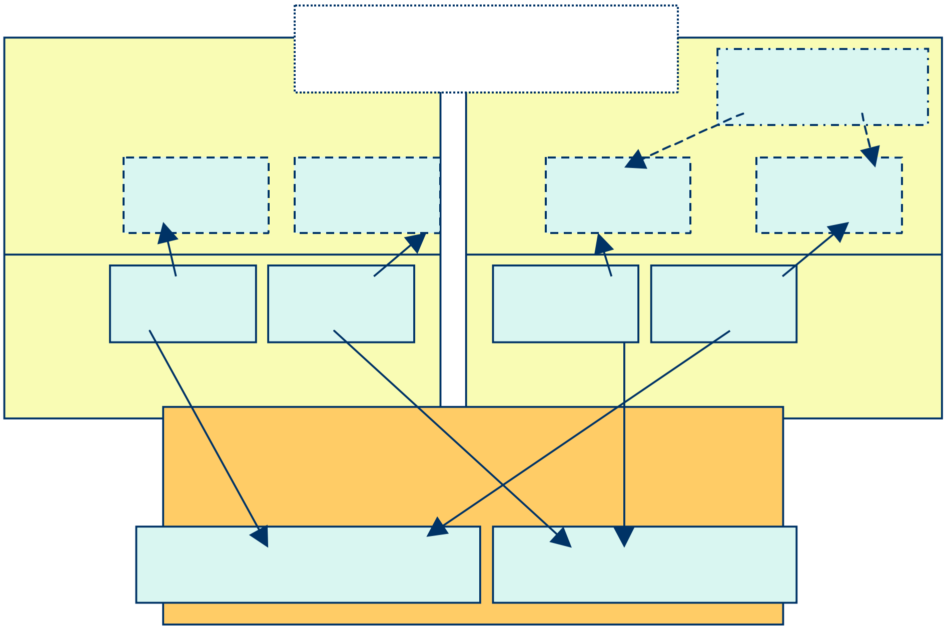


Figure 26: Crossed correspondence using Arguments and Slots

- synsets: the notion of synset is not the most suitable in a MT system, since each member of the synset can have a different syntactic and/or collocational behaviour in generation with respect to other members. Moreover, it is not possible to realize a cross-language *variant-to-variant* mapping by using the synset (this correspondence is feasible only between word senses). The multilingual extension of a monolingual wordnet-*like* lexicon is, however, important for a range of cross-languages applications, such as CLIR, CLIE and CRQA.

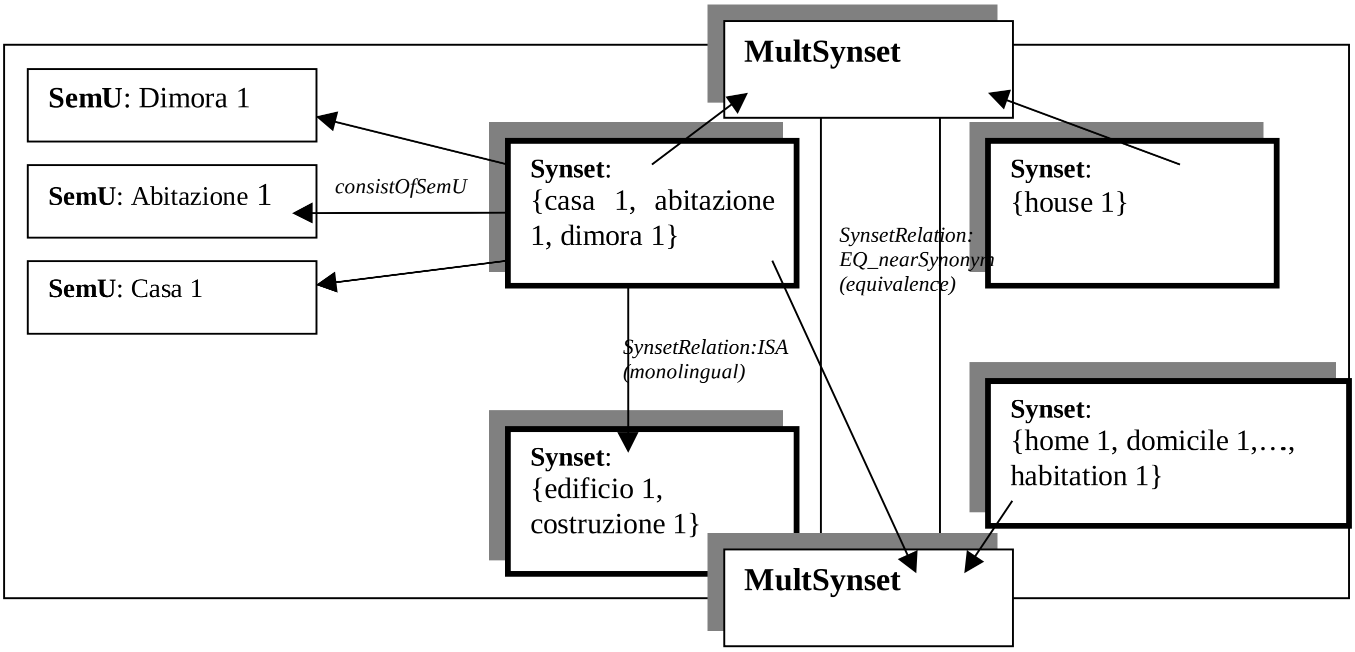

MultSynset
SemU: Dimora 1
SemU: Abitazione 1
SemU: Casa 1
consistOfSemU
Synset: {casa 1, abitazione 1, dimora 1}
Synset: {house 1}
SynsetRelation: EQ_nearSynonym (equivalence)
SynsetRelation:ISA (monolingual)
Synset: {edificio 1, costruzione 1}
Synset: {home 1, domicile 1,…, habitation 1}
MultSynset

Figure 27 : a possible scenario of correspondence between Synsets

6.2.6.4 TESTS AND ACTIONS MECHANISM

The core of the transfer is the mechanism of tests and actions of "*if…then*" type which apply respectively to source and target lexical objects. The table below shows the two main groups of operations, *constrain* and *add*, and the basic notions they work on.

Constrain	Add
Self	Slot
Slot	Syntagma
Syntagma	Syntactic Feature
Argument	Semantic Feature
	Argument
	Relation

Table 23

6.2.6.4.1 Constrain operations

They apply to source lexical objects (test operations) and to target lexical objects (action operations). By means of this family of operations it is possible to perform a restriction on the value of syntactic and semantic elements, forcing for example a slot of the syntactic frame to be realized by a certain phrase.

Constrain (Self)
The self can be constrained by adding syntactic or semantic features. Sub-elements of this operation are the operations consisting in adding semantic or syntactic features.

Constrain (Slot)
The slot of the syntactic frame can be constrained by changing its optionality status, by prohibiting it or by specifying the phrases filling it. This last operation implies an action of syntagma constraining.

Constrain (Syntagma)
A syntagma filling a position can be constrained with syntactic features and also by prohibiting its realization. The operation implied in this type of constraint is hence the addition of syntactic features.

Constrain (Argument)
The arguments of the semantic frame can be constrained specifying selectional restriction information.

6.2.6.4.2 "Add" Operations

They operate simply by adding the information individuated in the translation process to arrive to the correct equivalent.

Add (slot)
It allows to add a slot to a syntactic frame of the source or target lexical unit.

Add (argument)
It allows to add an argument to the semantic frame of the source or target lexical unit (cf. the constrain argument operation to impose selectional restriction on the new argument).

Add (syntagma)
It adds a new terminal or non terminal phrase to slots. It involves terminal and non terminal categories of the monolingual level.

Add (Syntactic Feature)
It allows to specify an auxiliary, a lexicalization of a phrase and to express morphosyntactic features: auxiliary, lexical and morphosyntactic features of the monolingual layer are involved at this level.

Add (Semantic Feature)
It allows to specify semantic information of the lexical unit or the arguments of its semantic frame. In multilingual perspective, it is useful to have the possibility to add all different types of information of semantic nature as introduced in paragraph 6.2.4.3: semantic type, domain and semantic features.

Add (Semantic Relation)
This operation applies in cases of correspondence between pairs of semantic units and can verify when, from the bilingual correspondence, the necessity of making a semantic relation explicit (needed only in cases of multilingual linking) emerges to better clarify the semantic purport.
It should be noted that all the operations work only at the level of multi-MILE, i.e. it is not the case to overload monolingual entries with information idiosyncratic of the multilingual layer and lexically-dependent on SL and TL language pairs.

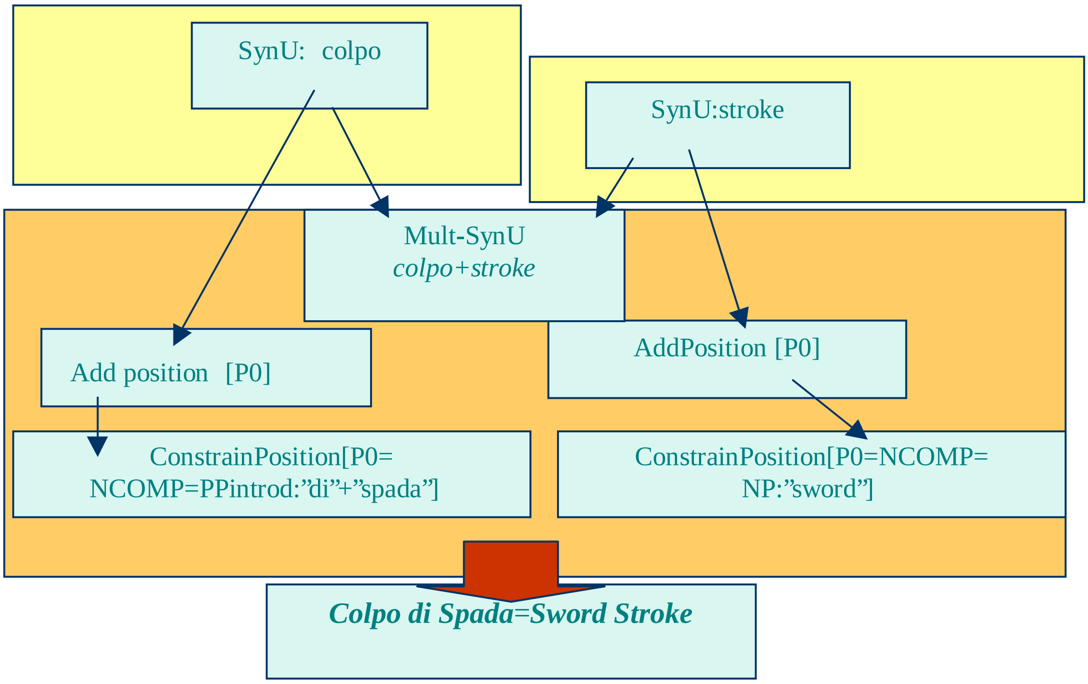


Figure 28: Using “constrain” and “add” operations to establish correspondences between multiwords

In 6.2.6.2 we introduced the necessity to stay open to approaches different from the "transfer-based" one. The MILE model offers the possibility to implement e.g. the interlingual approach to translation, by exploiting and specializing the semantic/conceptual layer of the monolingual level. The notion of semantic frame can be, indeed, extended to a more abstract notion of non-lexicalized predicate, where abstract primitives can be combined to realize a language independent, neutral and conceptual representation. In this sense, the representation resides outside the monolingual descriptions and does not need transfer rules, since the same internal representation is used for both the source and the target languages.

In Fig. 29 we give an example of a possible implementation of the interlingual approach: the interlingual predicate PredGIVE presents the prototypical argument structure of "giving" events where someone(agent) gives something(patient) to someone(beneficiary/2nd participant) irrespectively from any surface realization in actual languages. This intermediate predicative representation is pointed from the semantic frames of each monolingual module, thus allowing the engagement of the correct surface realization(s) in each language. As it can be seen, the two possible English realizations, the di-transitive and the prepositional dative alternation, are recovered via the set of interlingual and monolingual mapping rules between the semantic and syntactic frames.
This interlingual device has a further advantage of being exploitable in all the "giving events" (e.g. Eng. *to give*, *to donate*, It. *dare*, *regalare*, *elargire*, *donare* etc.)

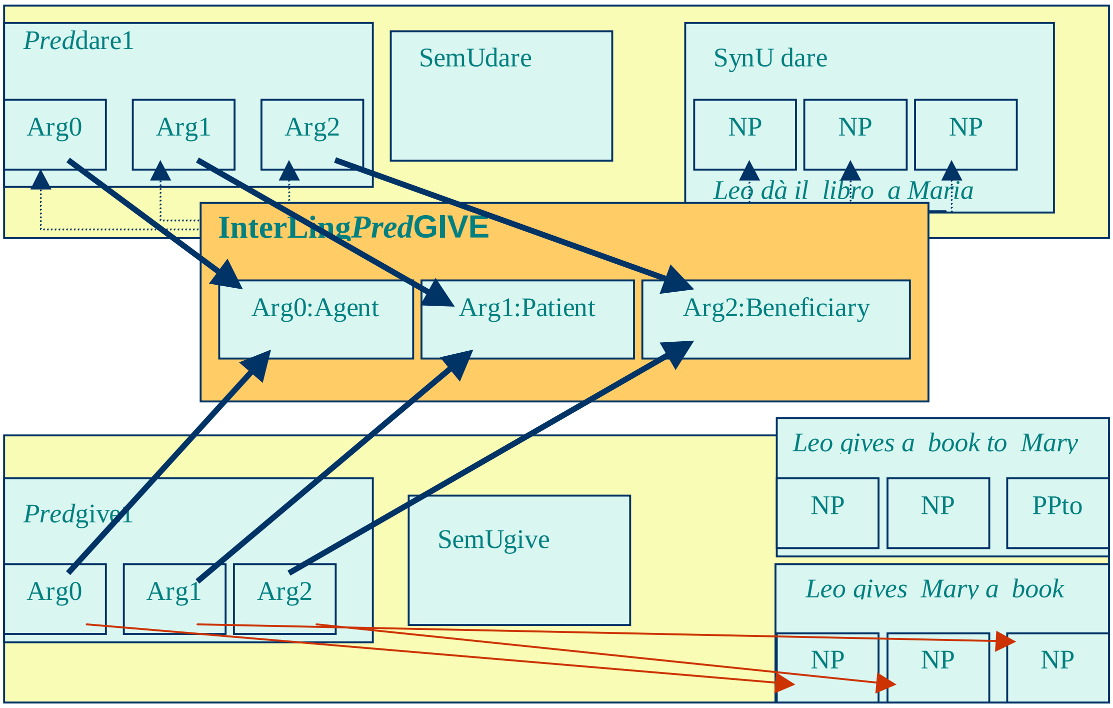


Figure 29: a possible implementation of the Interlingual approach in the MILE model

6.3 The MILE lexical model

6.3.1 The MILE architecture

In its general design MILE is envisaged as a highly *modular* and *layered* architecture (see Figure 30), as described in Calzolari *et al.* (2001b). Modularity concerns the *horizonta*" MILE organization, in which independent and yet linked modules target different dimensions of lexical entries. On the other hand, *at the "vertical" level*, a layered organization is necessary to allow for different degrees of granularity of lexical descriptions, so that both "shallow" and "deep" representations of lexical items can be captured. This feature is particularly *crucial in order to stay open to the different styles and approaches to the lexicon adopted by existing multilingual systems.*

At the top level, MILE includes two main modules, *mono-MILE*, providing monolingual lexical representations, and *multi-MILE*, where multilingual correspondences are defined. With this design choice the ISLE-CLWG intends also to address the particularly complex and yet *crucial issue of multilingual resource development through the integration of monolingual computational lexicons.*

Mono-MILE is organized into independent modules, respectively providing *morphological*, *syntactic* and *semantic* descriptions. The latter surely represents the core and the most challenging part of the ISLE-CLWG activities, together with the two other crucial topics of *collocations* and *multi-word expressions*, which have often remained outside standardization initiatives, and nevertheless have a crucial role at the multilingual level. This bias is motivated by the necessity of providing an answer to the most urgent needs and desiderata of next generation HLT, as also expressed by the industrial partners participating to the project. With respect to the issue of the representation of multi-word expressions in computational lexicons, the ISLE-CLWG has actively cooperated with the NSF sponsored XMELLT project (Calzolari *et al.*, 2002).

Multi-MILE specifies a formal environment for the characterization of multilingual correspondences between lexical items. In particular, source and target lexical entries can be linked by exploiting (possibly combined) aspects of their monolingual descriptions. Moreover, in multi-MILE both syntactic and semantic lexical representations can also be enriched, so as to achieve the granularity of lexical description required to establish proper multilingual correspondences, and which is possibly lacking in the original monolingual lexicons.

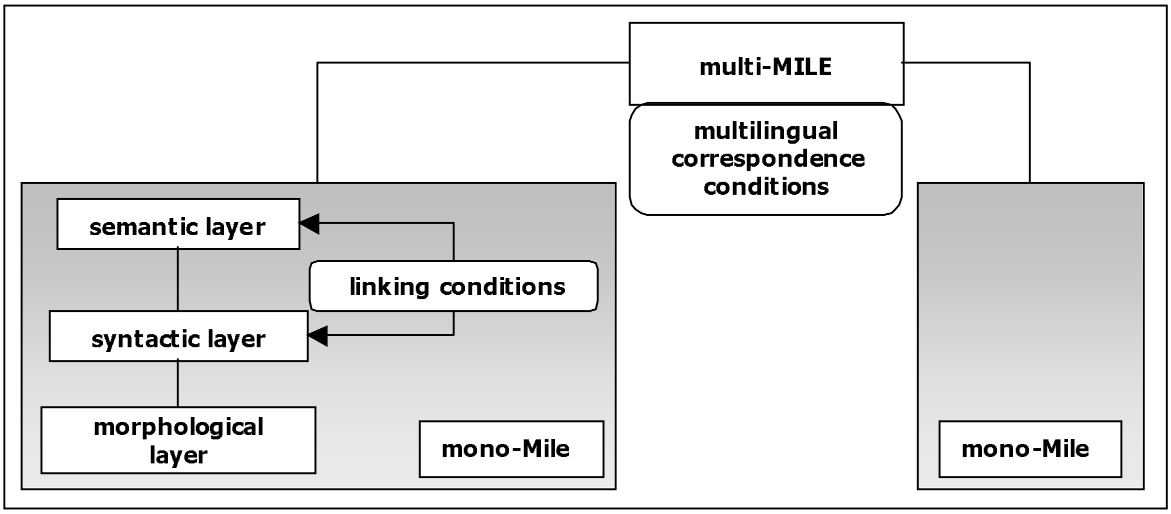


Figure 30

According to the ISLE approach, monolingual lexicons can thus be regarded as *pivot lexical repositories*, on top of which various language-to-language multilingual modules can be defined, where lexical correspondences are established by partly exploiting and partly enriching the monolingual descriptions. This architecture guarantees the independence of monolingual descriptions while allowing for the maximum degree of flexibility and consistency in reusing existing monolingual resources to build new bilingual lexicons.

The MILE architecture is intended to provide the common representational environment needed to implement such an approach to multilingual resource development, with the goal of maximizing the reuse, integration and extension of existing monolingual computational lexicons.

The following sections describe the MILE Lexical Model (MLM). This consists of an Entity-Relationship (E-R) diagram defining the entities of the lexical model and the way they can be combined to design an actual lexical entry. As such, the MLM does not correspond to a specific lexical entry, but is rather an *entry schema*, i.e. actually corresponding to a *lexical meta-entry*. This means that different possible lexical entries can be designed as instances of the schema provided by the MLM. Instance entries might therefore differ for the *type* of information they include (e.g. morphological, syntactic, semantic, monolingual or multilingual, etc.), and for the *depth* of lexical description.

The MLM includes two types of entities:

1. *MILE Lexical Classes* (MLC) - these represent the main building blocks of lexical entries. They formalize the basic lexical notions illustrated in 6.2. The MLM provides the definition of these classes, i.e. their attributes and the way they relate to each other (some complex classes are defined in terms of other classes). Classes represent notions like *syntactic feature*, *syntactic phrase*, *predicate*, *semantic relation*, *synset*, etc. The instances of MLCs are the *MILE Data Categories* (MDC). So for instance, NP and VP are data category instances of the class <Phrase>, and SUBJ and OBJ are data category instances of the class <Function>. Each MDC is identified by a URI. MDC can be either "user defined" or belong to "shared repositories".
2. *lexical operations* - these are special lexical entities which allow users to state complex conditions and perform complex operations over lexical entries. They will for instance allow lexicographers to establish multilingual conditions, link the slots within two different syntactic frames, link semantic arguments with syntactic slots, etc.

In order to distinguish the two types of lexical entities above in the E-R diagram, the name of MILE Lexical Classes is prefixed by MLC.

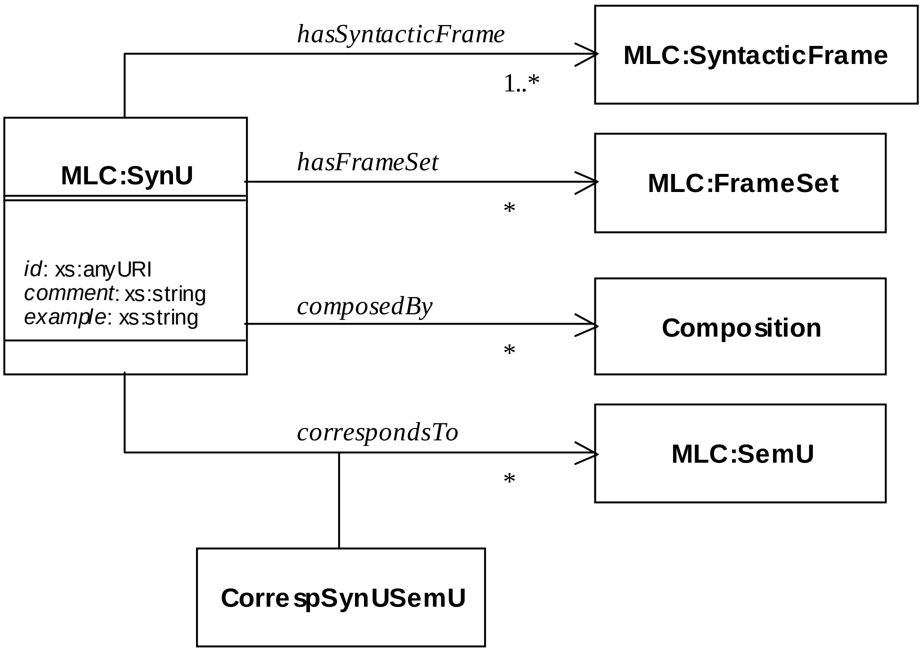
6.3.2 Syntactic layer

This layer includes the MLC (MILE Lexical Classes) corresponding to the syntactic basic lexical notions identified in section 6.2.3. Syntactic MLC formalize the notion of *syntactic subcategorization frame*. They directly rely on the specification of EAGLES syntax, integrated with further information types highly relevant for lexical description.

6.3.2.1 MLC:SYNU

A <SynU> (syntactic unit) is the class corresponding to a syntactic lexical entry. It is used to describe the syntactic subcategorization properties of lemmas, their possible associations with syntactic frames, etc.

Diagram



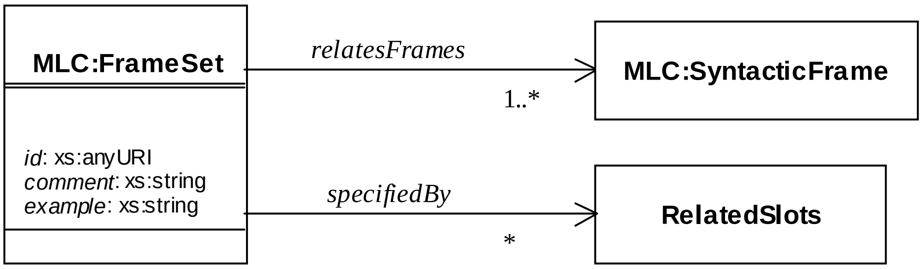
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SynU>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SynU>	xs:string	IMPLIED		
example	one or more example of the <SynU>	xs:string	IMPLIED		

6.3.2.2 MLC:FrameSet

A <FrameSet> is a MLC that expresses diathesis alternations of a lexical entry (e.g. causative-inchoative; dative alternation, etc.), by linking some of its syntactic frames. The syntactic slots of the linked frames can be also related via the <RelatedSlots> procedure.

Diagram



Attributes

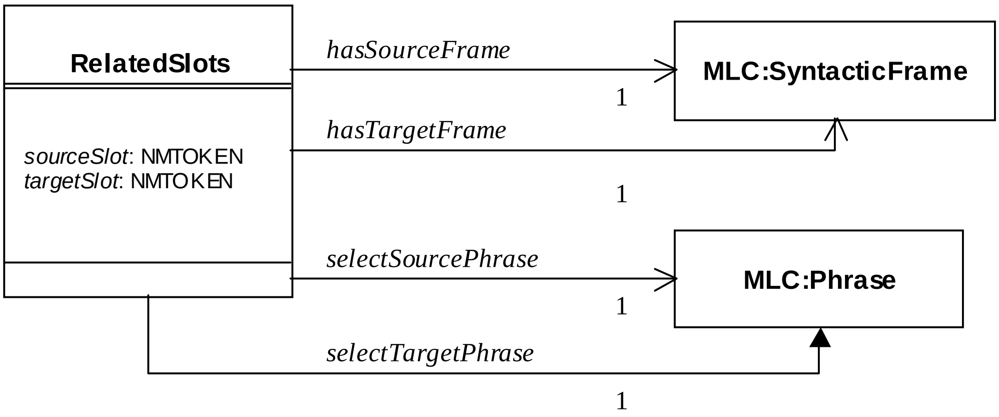
Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <FrameSet>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <FrameSet>	xs:string	IMPLIED		
example	one or more example of the	xs:string	IMPLIED		

	<FrameSet>				

6.3.2.3 RELATEDSLOTS

This entity formalizes the procedure of linking together two slots (and possibly the phrases realizing them) belonging to different syntactic frames.

Diagram



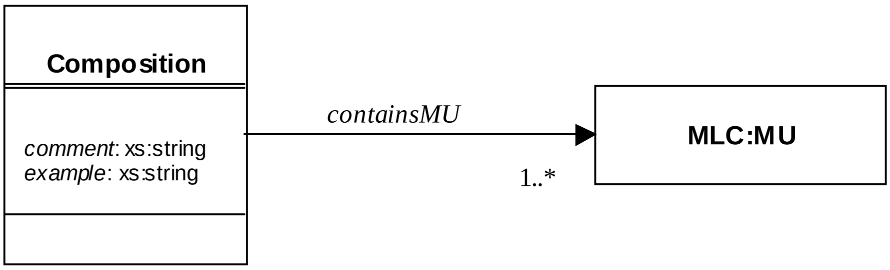
Attributes

Name	Description	Type	Use	Default	Fixed
sourceSlot	the number of the slot in the source description	NMTOKEN	REQUIRED		
targetSlot	the number of the slot in the source description	NMTOKEN	REQUIRED		

6.3.2.4 COMPOSITION

This entity encodes MWEs by simply listing their component lemmas (represented as <MU>).

Diagram



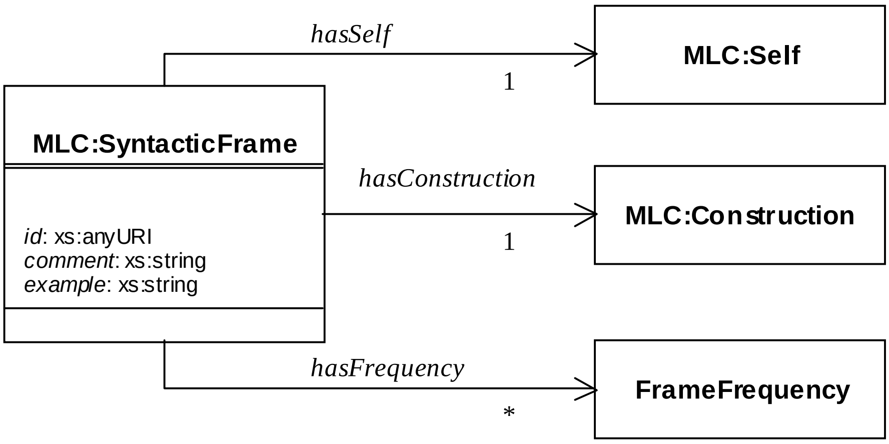
Attributes

Name	Description	Type	Use	Default	Fixed
comment	a comment or short description of the <Composition>	xs:string	IMPLIED		
example	one or more example of the <Composition>	xs:string	IMPLIED		

6.3.2.5 MLC:SyntacticFrame

The <SyntacticFrame> is the core class to specify subcategorization information. It is defined by: i.) the <Self> describing the properties of the head of the syntactic construction; ii.) the <Construction> specifying the syntactic arguments of the head; iii.) the <FrameFrequency>, which can be used to specify the frequency in a corpus of a certain subcategorization pattern.

Diagram



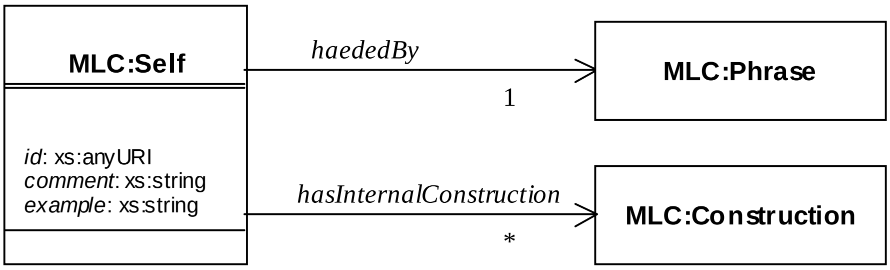
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SyntacticFrame>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SyntacticFrame>	xs:string	IMPLIED		
example	one or more example of the <SyntacticFrame>	xs:string	IMPLIED		

6.3.2.6 MLC:Self

The <Self> class specifies morphosyntactic constraints of the lexical unit being described. It can be *simple* or *complex*. Simple <Self> occurs when the lexical entry is not a MWE. Simple <Self> is only defined by a terminal phrase, specifying morphosyntactic properties of the lexical entry (e.g. syntactic category, auxiliary selection). Complex <Self> can be used to describe the internal syntactic structure of a MWE. The latter is expressed by including an internal <Construction> within the <Self>.

Diagram



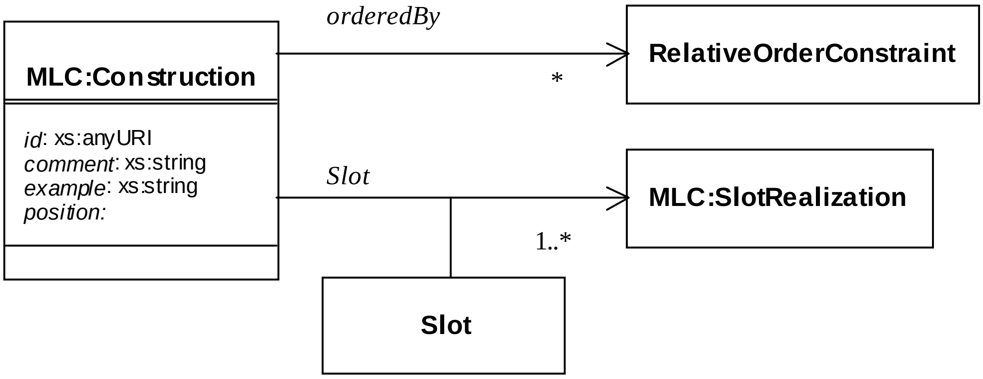
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <Self>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Self>	xs:string	IMPLIED		
example	one or more example of the <Self>	xs:string	IMPLIED		

6.3.2.7 MLC:CONSTRUCTION

The <Construction> class specifies the syntactic arguments of the entry. It consists of a number of syntactic slots which can be variously realized via <RelativeOrderConstraint>

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <Construction>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Construction>	xs:string	IMPLIED		
example	one or more example of the <Construction>	xs:string	IMPLIED		
position	whether the construction is internal or external to the <Self>	INTERNAL, EXTERNAL		EXTERNAL	

6.3.2.8 FRAMEFREQUENCY

The <FrameFrequency> specifies the frequency of a certain syntactic frame.

Attributes

Name	Description	Type	Use	Default	Fixed
comment	a comment or short description of the <FrameFrequency>	xs:string	IMPLIED		
corpus	the corpus with respect to which the frequency has been computed	xs:string	REQUIRED		

frequency	the frequency of the frame	NMTOKEN	REQUIRED		

6.3.2.9 RelativeOrderConstraint

This entity can be used to express constraints on the relative order of syntactic slots and their possible fillers.

Attributes

Name	Description	Type	Use	Default	Fixed
comment	a comment or short description of the <RelativeOrderConstraint>	xs:string	IMPLIED		
beforeSlot	the number of the slot occurring before	NMTOKEN	REQUIRED		
afterSlot	the number of the slot occurring after	NMTOKEN	REQUIRED		
beforePhrase	select a specific phrase (within the possible realizations of a slot) occurring before	IDREF	IMPLIED		
afterPhrase	select a specific phrase (within the possible realizations of a slot) occurring after	IDREF	IMPLIED		

6.3.2.10 Slot

The <Slot> corresponds to a syntactic slot within the subcategorization pattern described by the construction. Each slot is realized by syntactic phrases. Moreover a <Slot> can be either optional or obligatory.

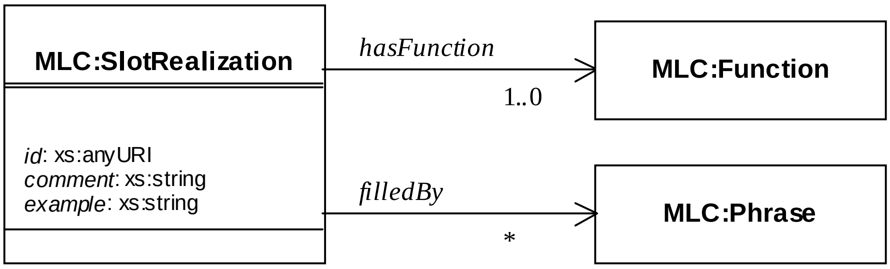
Attributes

Name	Description	Type	Use	Default	Fixed
slotId	a number identifying the slot	NMTOKEN	REQUIRED		
optional	whether the <slot> may be only implicitly realized or be necessarily present	YES, NO		YES	

6.3.2.11 MLC:SlotRealization

The <SlotRealization> class specifies the possible syntactic realizations of a slot within the construction. The <SlotRealization> consists in the specification of the slot grammatical function (e.g. subject, object) and of its possible syntactic fillers (phrases).

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SlotRealization>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SlotRealization>	xs:string	IMPLIED		
example	one or more example of the <SlotRealization>	xs:string	IMPLIED		

6.3.2.12 MLC:FUNCTION

The <Function> class specifies the grammatical function of syntactic slots (e.g. subject, direct object, etc.).

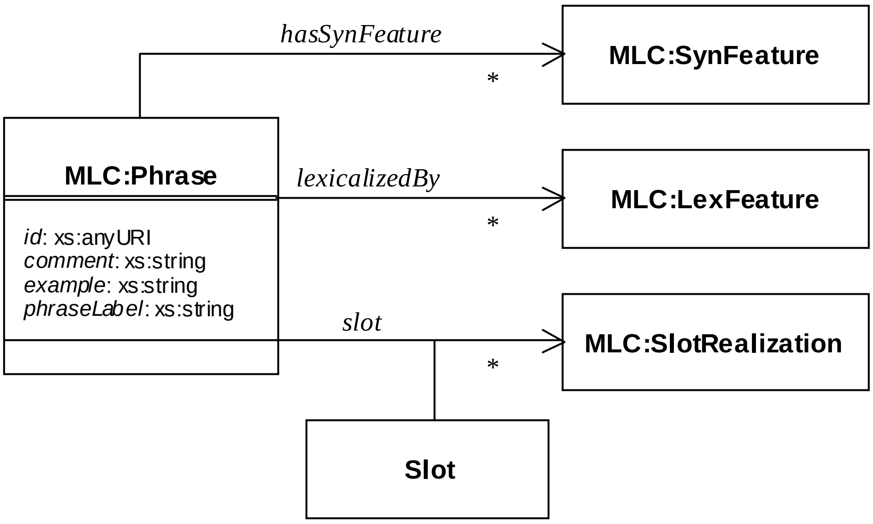
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <Function>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Function>	xs:string	IMPLIED		
example	one or more example of the <Function>	xs:string	IMPLIED		
functionName	the name of the grammatical function	xs:string	REQUIRED		

6.3.2.13 MLC:PHRASE

The <Phrase> class describes the phrases (terminal or non-terminal) realizing the slots in the construction and the self (e.g. NP, V, VP, etc.). Phrases are defined by bundles of *features*. Each phrase is identified by a label, specifying its category. Non-terminal phrases may be re-written, by specifying a series of slots. Phrases may also be partially or entirely lexicalized through <LexFeature> elements.

Diagram



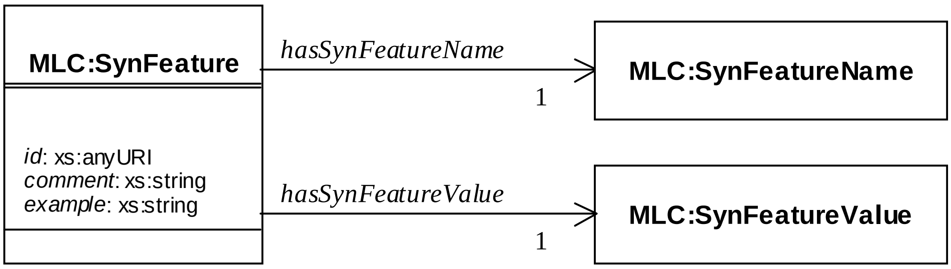
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <Phrase>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Phrase>	xs:string	IMPLIED		
example	one or more example of the <Phrase>	xs:string	IMPLIED		
phraseLabel	the syntactic label of the phrase	xs:string	REQUIRED		

6.3.2.14 MLC:SynFeature

This class specifies a (morpho)syntactic feature-value pair (e.g. **Gender** = *feminine*; **Tense** = *present*; **Control** = *subject_control*, etc.). These are used to build and describe phrases.

Diagram



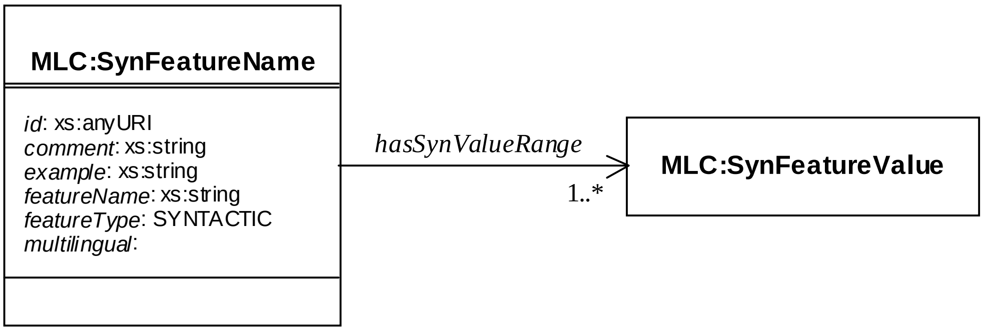
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SynFeature>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SynFeature>	xs:string	IMPLIED		
example	one or more example of the <SynFeature>	xs:string	IMPLIED		

6.3.2.15 MLC:SynFeatureName

This class specifies the (morpho)syntactic features (e.g. **Gender**, **Control**, **Tense**, **Number**, etc.) entering into the feature-value pairs. Features are defined by their range of values.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SynFeatureName>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SynFeatureName>	xs:string	IMPLIED		
example	one or more example of the <SynFeatureName>	xs:string	IMPLIED		
featureName	the name of the feature	xs:string	REQUIRED		
featureType	the type of the feature				SYNTACTIC
multilingual	whether the feature has monolingual or multilingual status	YES, NO		NO	

6.3.2.16 MLC:SynValue

It defines the possible values taken by features (e.g. *feminine*, *singular*, *present*, *subject_control*, etc.).

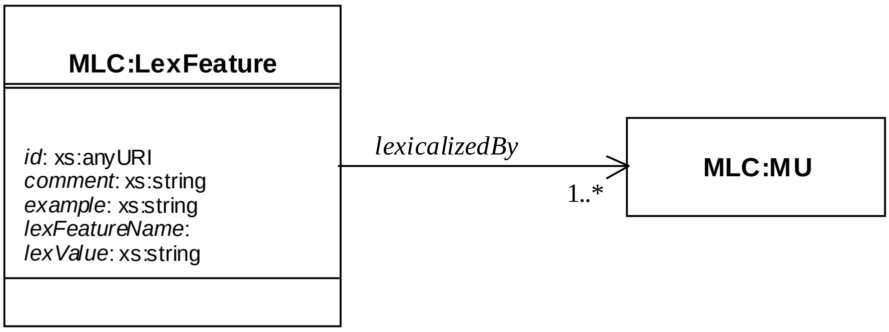
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SynValue>	xs:anyURI	REQUIRED		
valueName	the name of the feature value	xs:string	REQUIRED		

6.3.2.17 MLC:LexFeature

This class defines the possible patterns of lexicalizations of (parts of) syntactic phrases within a given lexicon. The lexicalization is expressed by pointing to the corresponding <MU>.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <LexFeature>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <LexFeature>	xs:string	IMPLIED		
example	one or more example of the <LexFeature>	xs:string	IMPLIED		
lexFeatureName	identifies the part of the phrase that is lexicalized	xs:string	REQUIRED		
lexValue	the canonical form of the lexicalizing <MU>	xs:string	REQUIRED		

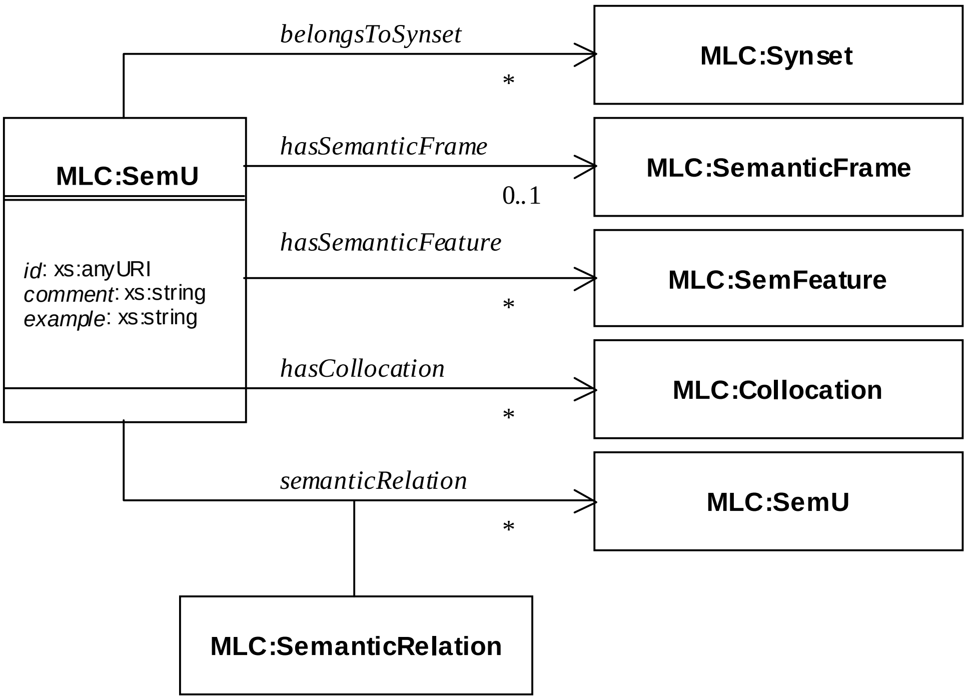
6.3.3 Semantic layer

This layer includes the MLC (MILE Lexical Classes) corresponding to the semantic basic lexical notions identified in section 6.2.4.

6.3.3.1 MLC:SEMU

A <SemU> (Semantic Unit) describes the meaning of a morphological unit. Each lemma may have more than one <SemU>. The <SemU> concentrates the semantic information corresponding to the sense of a lexical entry. Semantic information can consist of: i.) semantic features of different types (domain, ontology, etc.); ii.) the synsets to which the SemU belongs; iv.) semantic relations with other <SemU>; v.) a semantic frame specifying a list of semantic arguments; vi.) a set of possible collocations.

Diagram



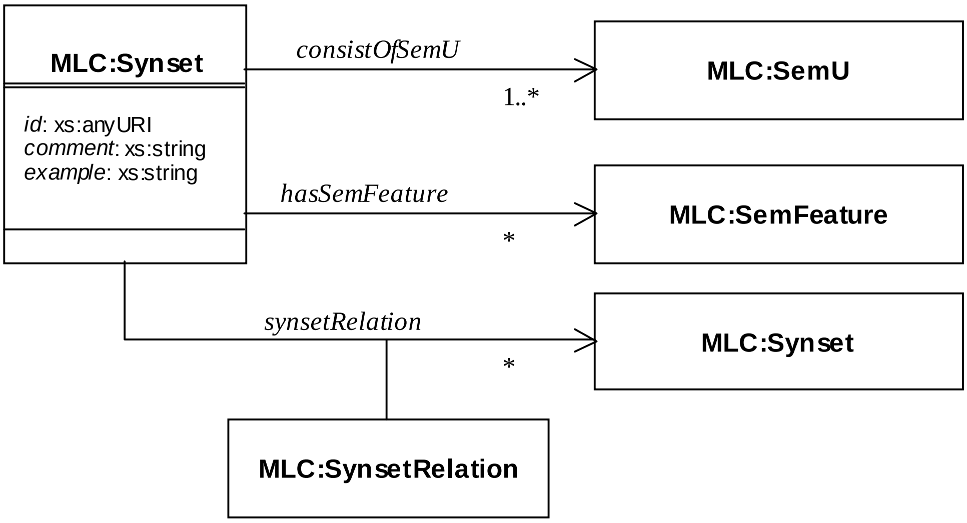
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SemU>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SemU>	xs:string	IMPLIED		
example	one or more example of the <SemU>	xs:string	IMPLIED		

6.3.3.2 MLC:SYNSET

This class formalizes the notion of *synset* as defined in WordNet (Fellbaum 1998). A synset is a set of synonyms and can be related to other synsets.

Diagram



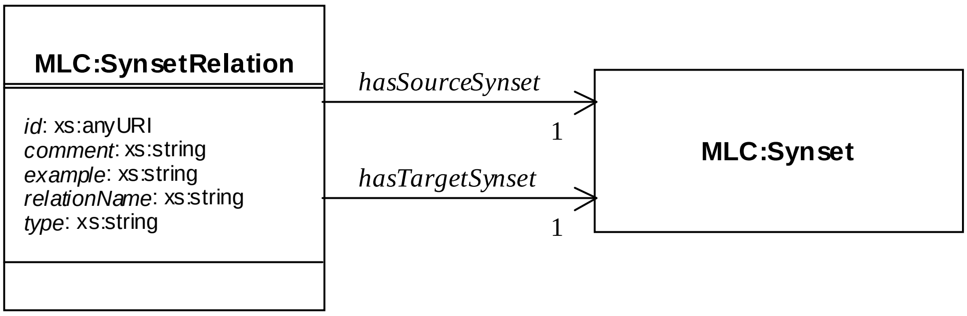
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <Synset>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Synset>	xs:string	IMPLIED		
example	one or more example of the <Synset>	xs:string	IMPLIED		

6.3.3.3 MLC:SynsetRelation

This class defines relations connecting synsets, as specified in Wordnet (e.g. hyperonymy, meronymy, etc.)

Diagram



Attributes

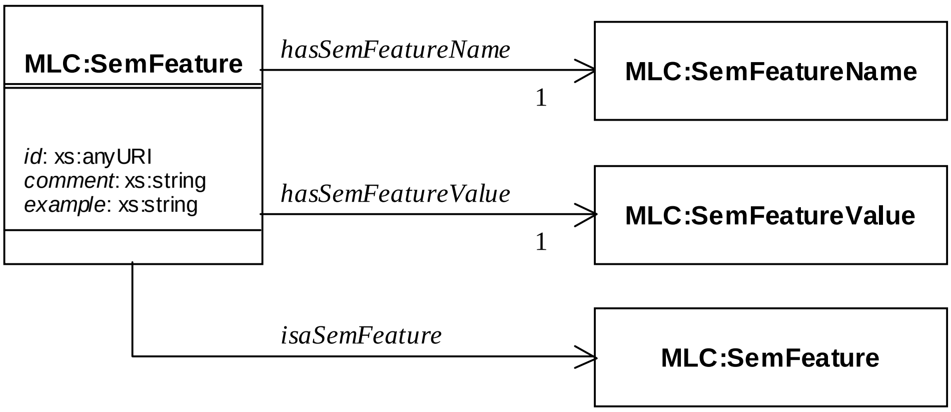
Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SynsetRelation>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SynsetRelation>	xs:string	IMPLIED		
example	one or more example of the	xs:string	IMPLIED		

	<SynsetRelation>				
relationName	the name of the <SynsetRelation>	xs:string	REQUIRED		
type	the type of the relation (e.g. monolingual, thematic, etc.)	xs:string	IMPLIED		

6.3.3.4 MLC:SemFeature

This class specifies a semantic feature-value pair (e.g. **Domain** = *medicine*; **Human** = *yes*; **SemanticType** = *group*, etc.). Semantic features are used to describe <SemU>, <Synset> or to specify selectional preferences on the semantic arguments. Semantic features can also be hierarchically structured.

Diagram



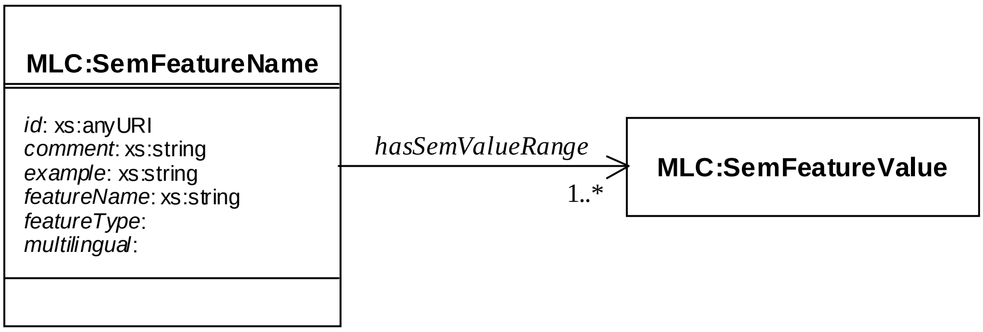
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SemFeature>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SemFeature>	xs:string	IMPLIED		
example	one or more example of the <SemFeature>	xs:string	IMPLIED		

6.3.3.5 MLC:SemFeatureName

This class specifies the semantic features (e.g. **SemanticType**, **Domain**, etc.) entering into the semantic feature-value pairs. Features are defined by their range of values.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SemFeatureName>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SemFeatureName>	xs:string	IMPLIED		
example	one or more example of the <SemFeatureName>	xs:string	IMPLIED		
featureName	the name of the feature	xs:string	REQUIRED		
featureType	the type of the feature	DOMAIN, ONTOLOGY, ASPECTUAL, STYLISTIC, PRAGMATIC, QUALIA, RESTRICTIVE	REQUIRED		
multilingual	whether the feature has monolingual or multilingual status	YES, NO		NO	

6.3.3.6 MLC:SEMVALUE

It defines the possible values taken by features (e.g. *group*, *medicine*, *animate*, etc.).

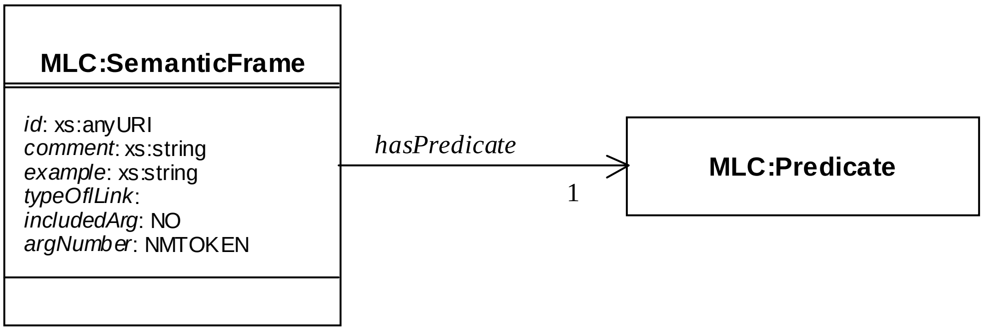
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SemValue>	xs:anyURI	REQUIRED		
valueName	the name of the feature value	xs:string	REQUIRED		

6.3.3.7 MLC:SEMANTICFRAME

This class defines the semantic frame of a <SemU>. Semantic frames specify the predicative argument structure of a lexical entry. The <SemanticFrame> is described in terms of a predicate and the type of link between the <SemU> and the predicate (the predicate is in turn defined in terms of the number and types of its arguments). Different <SemU> (possibly of words belonging to different parts of speech) may share the same predicate in the predicative representation. For instance, the verb *destroy* and the nouns *destruction* and *destroyer* may be represented as all sharing the same predicate DESTROY. The same holds for the verb *employ*, and the nouns *employment*, *employer* and *employee*, which can share the same predicate EMPLOY. These <SemU> however differ for the type of relation they have with this predicate. This difference is expressed by the attribute *typeOfLink* in the <SemanticFrame> class. The recommended values of these attribute are: MASTER (for verbs, relational nouns, representations, amounts, nouns with support verbs, etc.), VERBNOM (for *nomina actionis*; e.g. *destruction*), AGENTNOM (for *nomina agentis*; e.g. *destroyer*), PATIENTNOM (for object nominalizations; e.g. *employee*), ADJNOM (for deadjectival nouns; e.g. *patience*).

Diagram



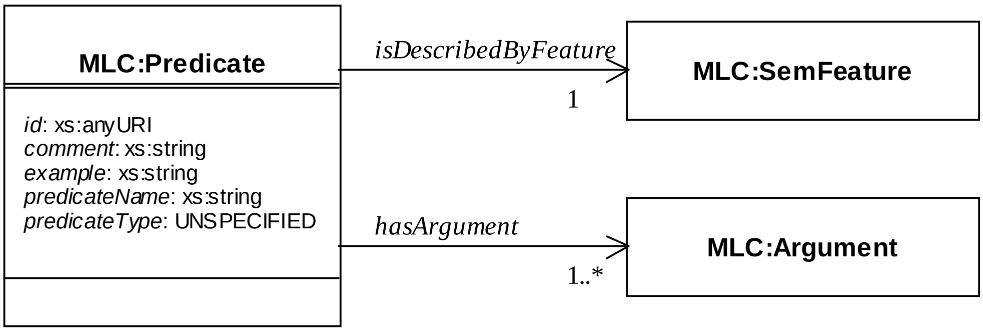
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SemanticFrame>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SemanticFrame>	xs:string	IMPLIED		
example	one or more example of the <SemanticFrame>	xs:string	IMPLIED		
typeOfLink	whether the <SemU>has a master relation (for verbs, relational nouns, representations, amounts, nouns with support verbs, etc.) with a predicate or not, i.e. whether it is the privileged and most neutral lexicalization of that predicate	MASTER, VERBNOM, AGENTNOM, PATIENTNOM, ADJNOM	REQUIRED		
includedArg	whether the <SemU> lexically absorbs one of the arguments of the predicate. The absorbed argument is thus not linked to the syntax.	YES, NO		NO	
argNumber	the number of the absorbed argument	NMTOKEN	IMPLIED		

6.3.3.8 MLC:PREDICATE

This class defines the predicates entering into the <Semantic Frame>. Predicates can be monolingual or multilingual. Multilingual predicates can be used to define ‘interlingua”-like semantic representations. <Predicate> is specified by the number and types of its arguments and can be further described by semantic features.

Diagram



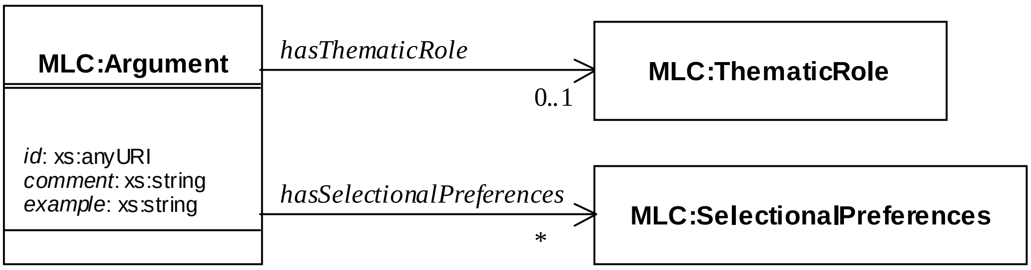
Attributes

Name	Description	Type	Use	Default	Fixed
Id	a unique identifier of the <Predicate>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Predicate>	xs:string	IMPLIED		
Example	one or more example of the <Predicate>	xs:string	IMPLIED		
predicateName	the name of the <Predicate>	xs:string	REQUIRED		
predicateType	the type of the predicate	PRIMITIVE, LEXICAL, UNSPECIFIED		UNSPECIFIED	

6.3.3.9 MLC:ARGUMENT

This class defines the arguments entering into the specification of a predicate. Each <Argument> can be characterized by a thematic (or semantic) role and/or by selectional preferences.

Diagram



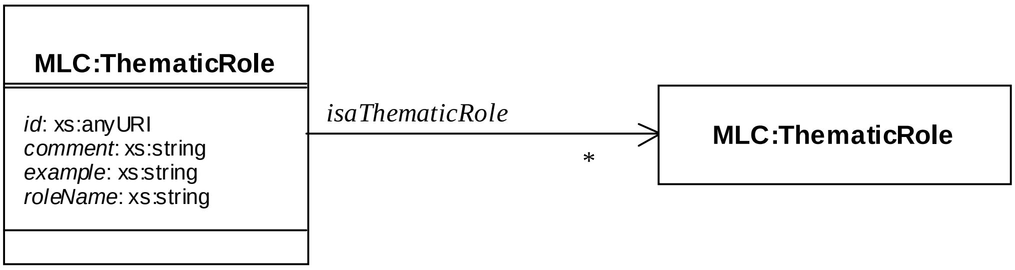
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <Argument>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Argument>	xs:string	IMPLIED		
example	one or more example of the <Argument>	xs:string	IMPLIED		

6.3.3.10 MLC:ThematicRole

This class defines the thematic (or semantic roles) that can be used to specify the arguments within a semantic frames. Possible instances of this class are *Agent, Patient, Experiencer,* etc. Thematic Roles can be hierarchically organized.

Diagram



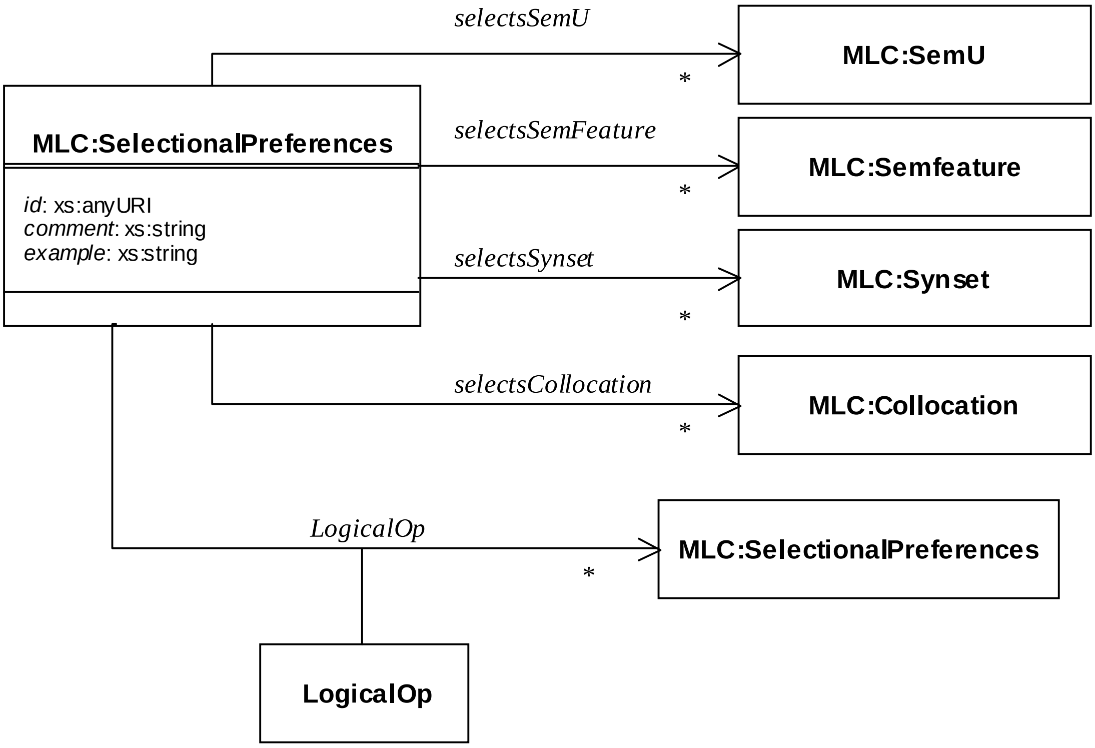
Attri butes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <ThematicRole>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <ThematicRole>	xs:string	IMPLIED		
example	one or more example of the <ThematicRole>	xs:string	IMPLIED		
roleName	the name of the < ThematicRole>	xs:string	REQUIRED		

6.3.3.11 MLC:SelectionalPreferences

This class defines the selectional preferences of semantic frame arguments. Selectional preferences is a cluster of information that semantically constrain the possible realizations of the semantic frame arguments. This cluster may include: i.) semantic features, ii.) synsets, iii.) collocations, iv.) particular semantic units, and v) a combination of all these types of lexical information. Moreover, it is possible to express ‘logically” complex selectional preferences, i.e. to combine various selectional preferences with logical operators.

Diagram

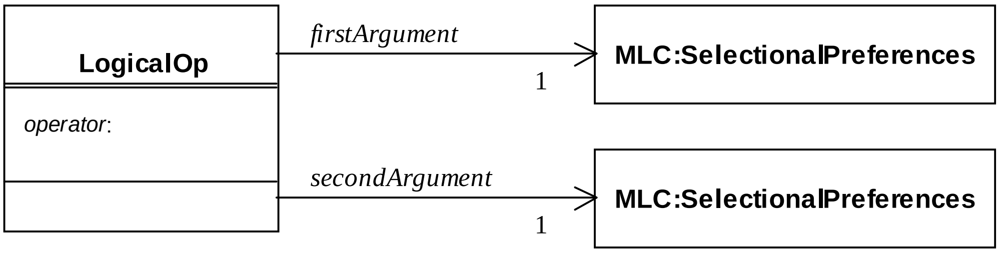


Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SelectionalPreferences>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SelectionalPreferences>	xs:string	IMPLIED		
example	one or more example of the <SelectioanlPreferences>	xs:string	IMPLIED		

6.3.3.12 LOGICALOP

This entity can be used to express logical combinations of lexical objects: selectional preferences, etc.



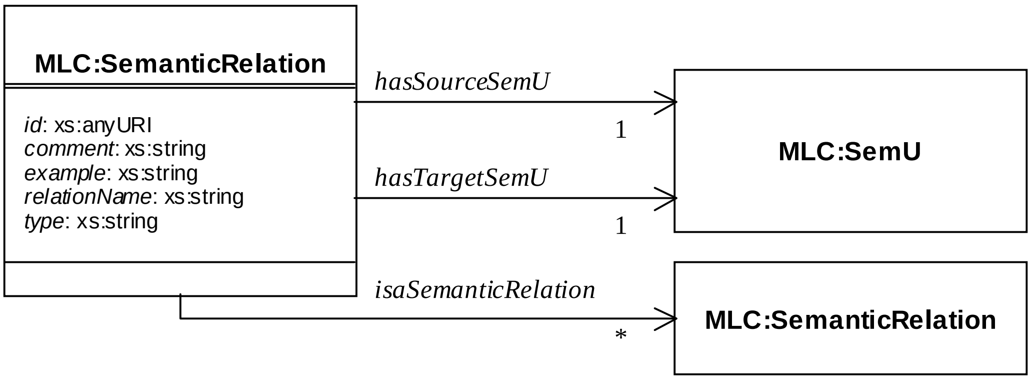
Attributes

Name	Description	Type	Use	Default	Fixed
operator	the name of the logical operator	AND, OR	REQUIRED		

6.3.3.13 MLC:SEMANTICRELATION

This class defines semantic relations linking two <SemU>. Possible instances of this class are *hyponymy*, *meronymy*, etc. While <SynsetRelation> links two synsets, i.e. sets of synonyms, <SemanticRelation> specifies the semantic content of a source <SemU> by linking it to another target <SemU>

Diagram



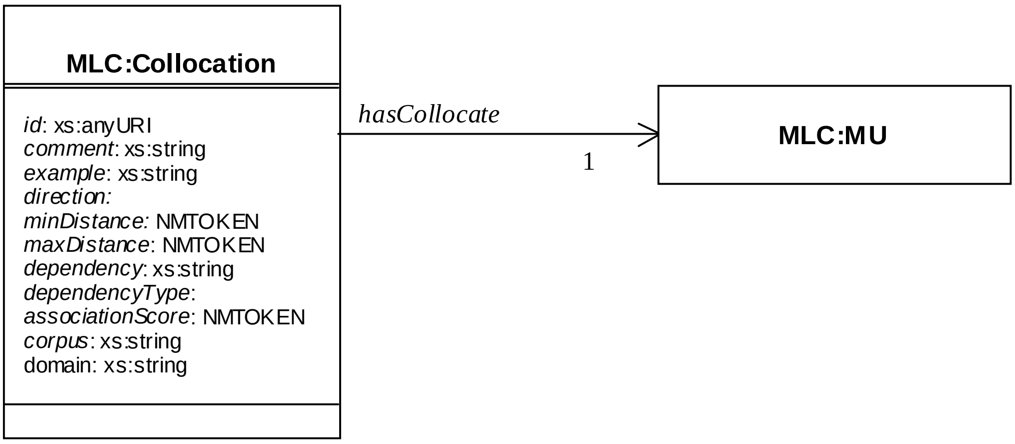
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SemanticRelation>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SemanticRelation>	xs:string	IMPLIED		
example	one or more example of the <SemanticRelation>	xs:string	IMPLIED		
relationName	the name of the <SemanticRelation>	xs:string	REQUIRED		
type	the type of the relation (e.g. monolingual, thematic, etc.)	xs:string	IMPLIED		

6.3.3.14 MLC:COLLOCATION

This class can be used to specify the collocations of the lexical entry. The semantic content of a lexical entry (i.e. one of its <SemUs>) can thus be characterized in terms of its collocations, intended as the word co-occurrence relations in which it appears in texts (Sinclair 1991). The <Collocation> class consists of a relation with a <MU>, the latter representing the collocate word. The attributes of the <Collocation> class are consistent with the specifications proposed in the *EAGLES Recommendations on Lexical Semantics* (Sanfilippo, 1999). In particular, the dependencyType attribute gives information about the dependency configuration being described, in particular about the relationship between the word sense entry and the collocate word. Four dependency configuration have been identified: i.) h2d (head to dependant), ii) d2h (dependant to head), iii.) d2d (dependant to dependant), and iv.) h2h (head to head). For details on these relations, cf. Sanfilippo 1999.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <Collocation>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <Collocation>	xs:string	IMPLIED		
example	one or more example of the <Collocation>	xs:string	IMPLIED		
direction	specifies the right or left location of the collocates with respect to the word being defined	RIGHT, LEFT	REQUIRED		
minDistance	the minimal distance between the co-occurring words	NMTOKEN	IMPLIED		
maxDistance	the maximal distance between the co-occurring words	NMTOKEN	IMPLIED		
dependency	the grammatical function of the collocate with respect to the head (e.g. subj, obj, comp, etc.)	xs:string	REQUIRED		
dependencyType	the relationship holding between the word sense entry and the collocate word	H2D, D2H, D2D, H2H	IMPLIED		
associationScore	the strength of the association with the co-occurring word (e.g. mutual information)	NMTOKEN	IMPLIED		
corpus	the corpus from which the collocation has been extracted	xs:string	IMPLIED		
domain	the domain of the collocation	xs: string	IMPLIED		

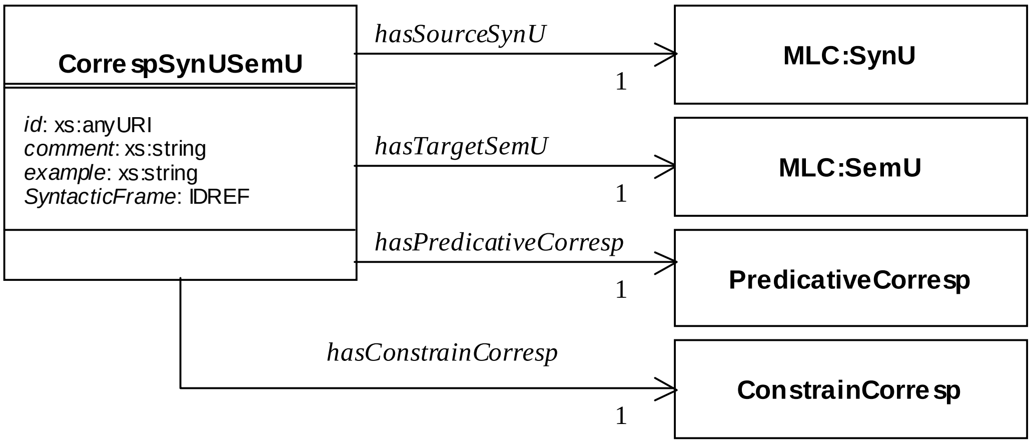
6.3.4 Syntax-Semantics Linking

This layer specifies the lexical objects necessary to link syntactic and semantic entries.

6.3.4.1 CORRESPSYNUSEMU

This object links a <SynU> to a <SemU>. It is also possible to link syntactic slots to semantic arguments, ands to specify constraints on the <SynU>-<SemU> association

Diagram



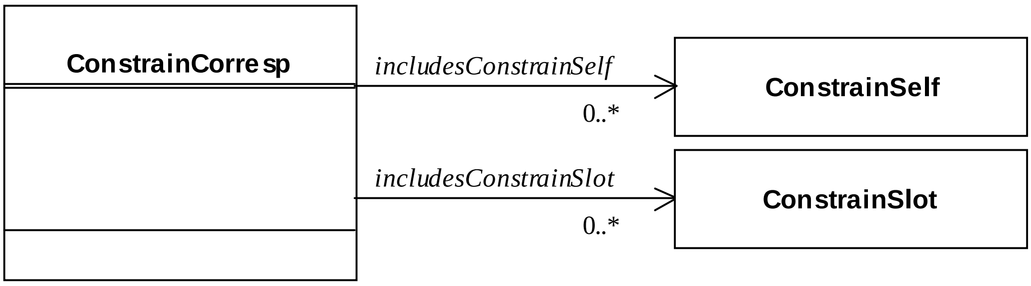
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <CorrespSynUSemU>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <CorrespSynUSemU>	xs:string	IMPLIED		
example	one or more example of the <CorrespSynUSemU>	xs:string	IMPLIED		
SyntacticFrame	the identifier of the <SyntacticFrame>, in the case in which the link to the <SemU>holds only for a particular syntactic frame of the <SynU>	IDREF	IMPLIED		

6.3.4.2 CONSTRAINCORRESP

This objects contains the constrains to the syntax-semantics association (for the definition of <ConstrainSelf> and <ConstrainSlot>, see below in the multilingual layer)

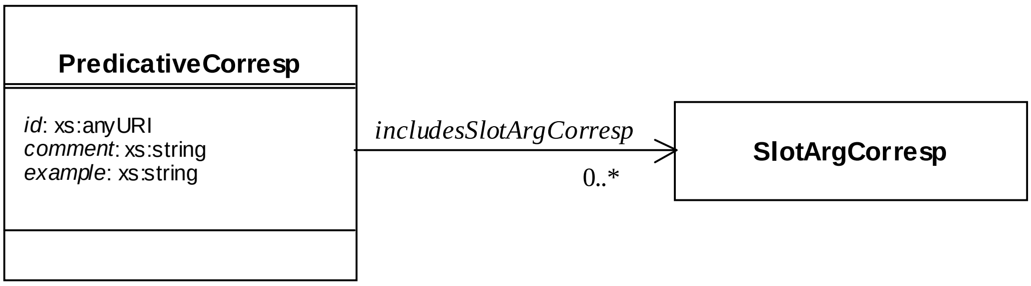
Diagram



6.3.4.3 PREDICATIVECORRESP

This object contains the associations between the specific syntactic positions and semantic arguments.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <PredicativeCorresp>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <PredicativeCorresp>	xs:string	IMPLIED		
example	one or more example of the <PredicativeCorresp>	xs:string	IMPLIED		

6.3.4.4 SLOTARGCORRESP

This object links a syntactic position to a semantic argument

Attributes

Name	Description	Type	Use	Default	Fixed
slotNumber	the number of the syntactic <Slot>	NMTOKEN	REQUIRED		
argNumber	the number of the semantic <Argument>	NMTOKEN	REQUIRED		
slotPosition	whether the linked position is external to the <Self> or internal to the <Self>. This	EXTERNAL, INTERNAL		EXTERNAL	

	attribute can be used to eventually link an argument to a sub-part of a MWE	

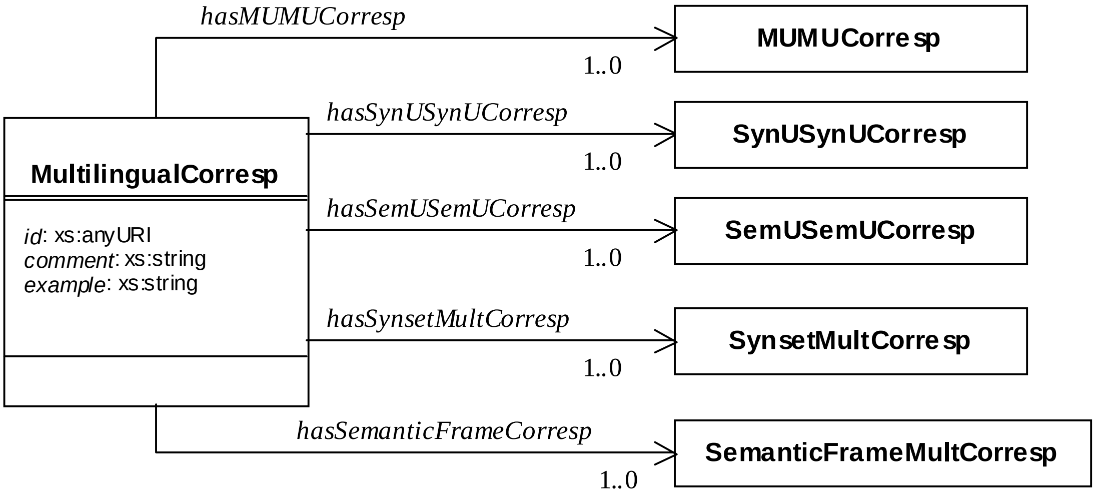
6.3.5 Multilingual layer

This layer includes the MLC and lexical objects corresponding to the multilingual correspondences and operations identified in section 6.2.5.

6.3.5.1 MULTILINGUALCORRESP

This object expresses the multilingual correspondences between lexical entries. These can consist of: 1. link between <MU>; 2. link between <SynU>; 3. link between <SemU>; 4. link between a <Synset> and a <MultilingualSynset>; 5. link between a <SemU> and a <MultilingualSemanticFrame>. (1)-(3) implement the transfer-based approach to multlinguality, while (4) and (5) implement the interlingua approach.

Diagram



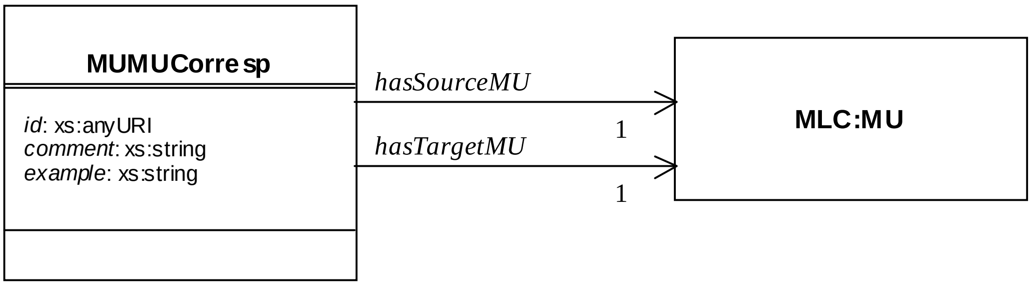
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <MultilingualCorresp>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <MultilingualCorresp>	xs:string	IMPLIED		
example	one or more example of the <MultilingualCorresp>	xs:string	IMPLIED		

6.3.5.2 MUMUCORRESP

This object expresses a link between a source and a target <MU>.

Diagram



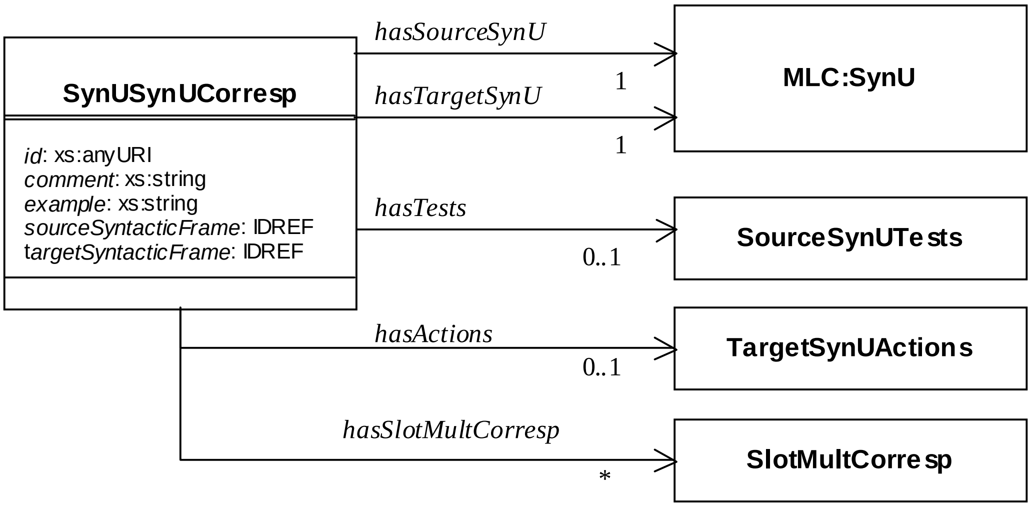
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <MUMUCorresp>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <MUMUCorresp>	xs:string	IMPLIED		
example	one or more example of the <MUMUCorresp>	xs:string	IMPLIED		

6.3.5.3 SYNUSYNUCORRESP

This object expresses a link between a source and a target SynU. Complex conditions on the source (tests) and on the target (actions) <SynU> can also be represented. It is also possible to link the syntactic slots of the two <SynU>.

Diagram



Attributes

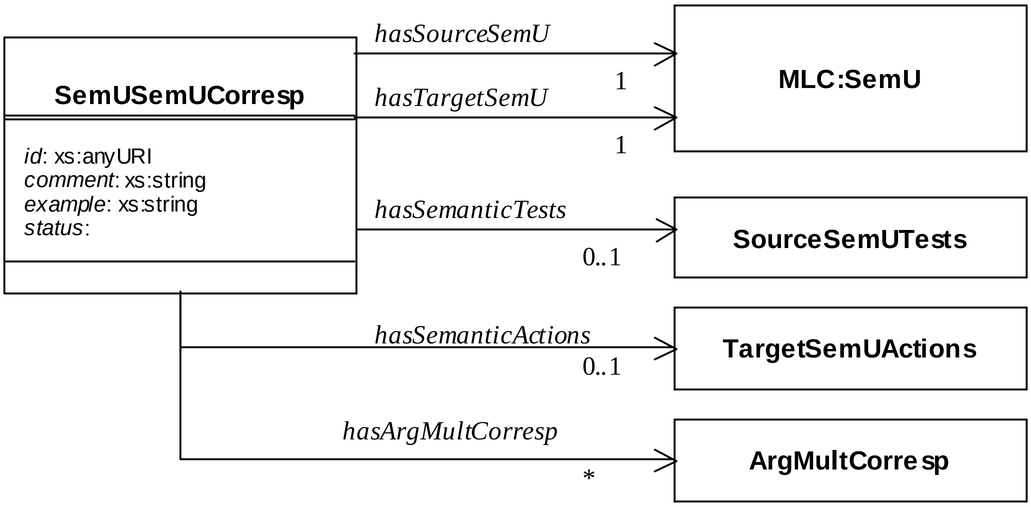
Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SynUSynUCorresp>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SynUSynUCorresp>	xs:string	IMPLIED		

example	one or more example of the <SynUSynUCorresp>	xs:string	IMPLIED		
sourceSyntacticFrame	the identifier of the source<SyntacticFrame>, in the case in which the multilingual link holds only of a particular <SyntacticFrame> of the source <SynU>	IDREF	IMPLIED		
targetSyntacticFrame	the identifier of the target<SyntacticFrame>, in the case in which the multilingual link holds only of a particular <SyntacticFrame> of the target <SynU>	IDREF	IMPLIED		

6.3.5.4 SemUSemUCorresp

This object expresses a link between a source and a target SemU. Complex conditions on the source (tests) and on the target (actions) <SemU> can also be represented. It is also possible to link the arguments in the <SemanticFrame> of the two <SemU>.

Diagram



Attributes

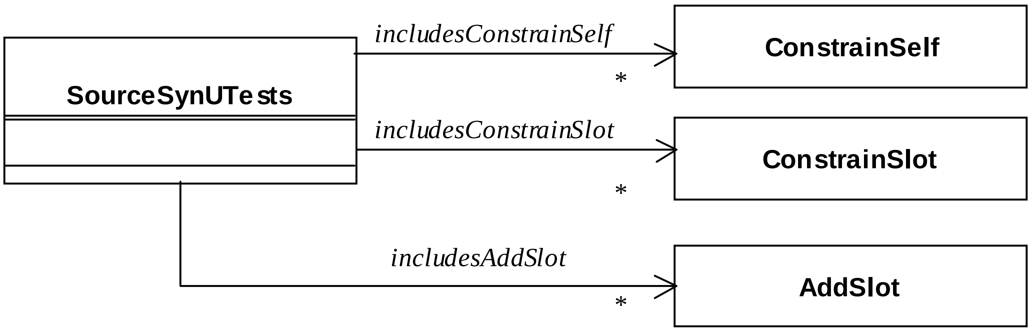
Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SynUSynUCorresp>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SynUSynUCorresp>	xs:string	IMPLIED		
example	one or more example of	xs:string	IMPLIED		

	the <SynUSynUCorresp>	
status	whether the linked elements are fully or partially equivalent	FULLEQUIVALENT, IMPLIED PARTIALEQUIVALENT

6.3.5.5 SourceSynUTests

This lexical object contains the syntactic tests and constraints on the source <SynU>

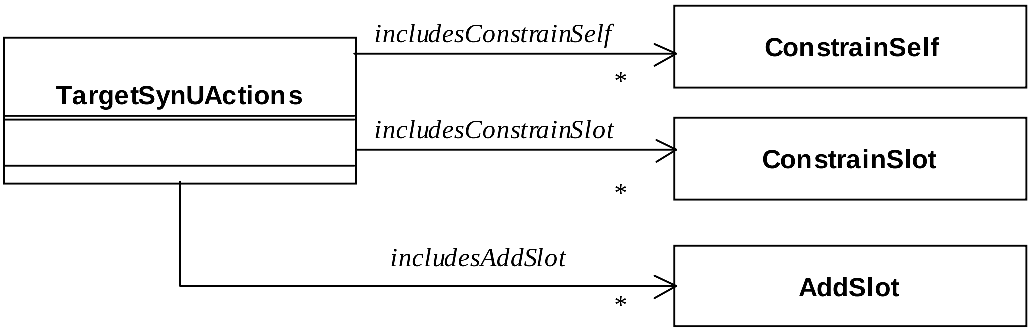
Diagram



6.3.5.6 TargetSynUActions

This lexical object contains the syntactic actions on the target <SynU>

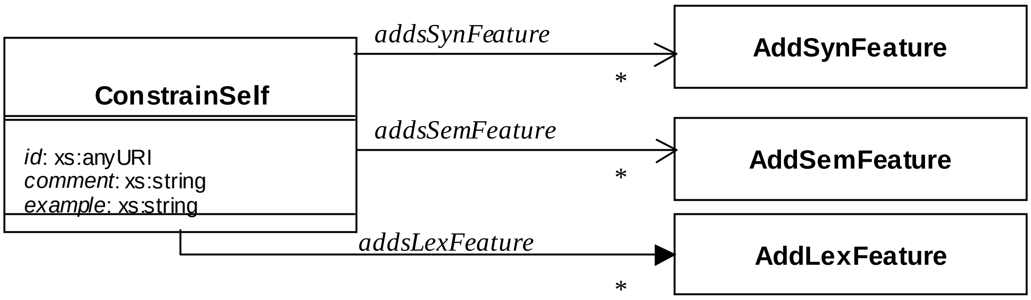
Diagram



6.3.5.7 ConstrainSelf

This object constrains the realization of the <Self> of a <SyntacticFrame>, by adding a syntactic or a semantic feature.

Diagram



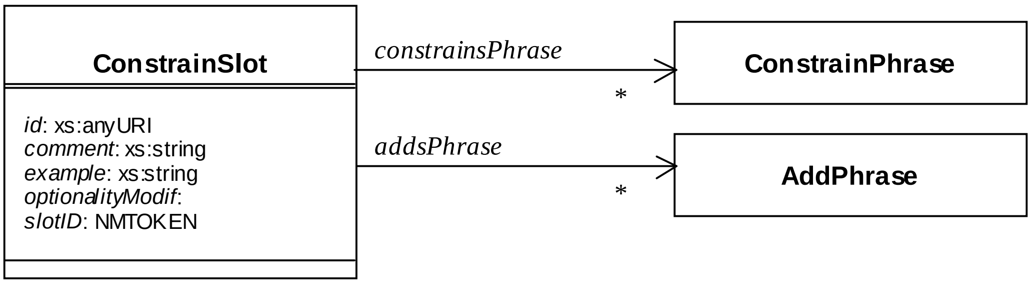
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <ConstrainSelf>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <ConstrainSelf>	xs:string	IMPLIED		
example	one or more example of the <ConstrainSelf>	xs:string	IMPLIED		

6.3.5.8 CONSTRAINSLOT

A slot can be constrained either by changing its optional attribute, or by specifying the phrases that fill it (<ConstrainPhrase>, <AddPhrase>)

Diagram



Attributes

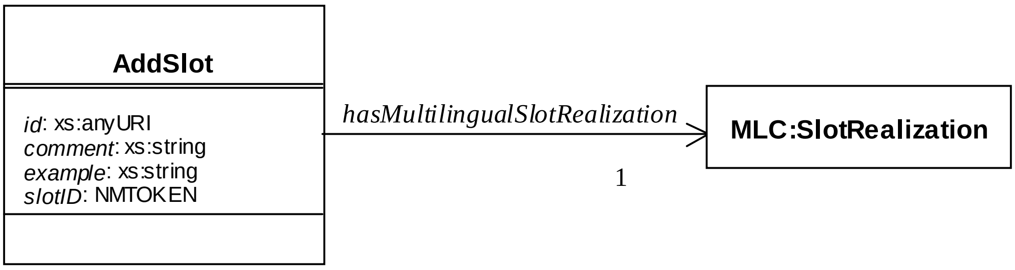
Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <ConstrainSlot>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <ConstrainSlot>	xs:string	IMPLIED		
example	one or more example of the <ConstrainSlot>	xs:string	IMPLIED		
slotID	the number of the slot to be constrained	NMTOKEN	REQUIRED		

optionalityModif	whether the optional status of the slot changes or not	COMPULSORY, PROHIBITED, NOMODIF		NOMODIF	

6.3.5.9 ADDSLOT

This object adds a new syntactic <Slot> to an existing <SyntacticFrame>

Diagram



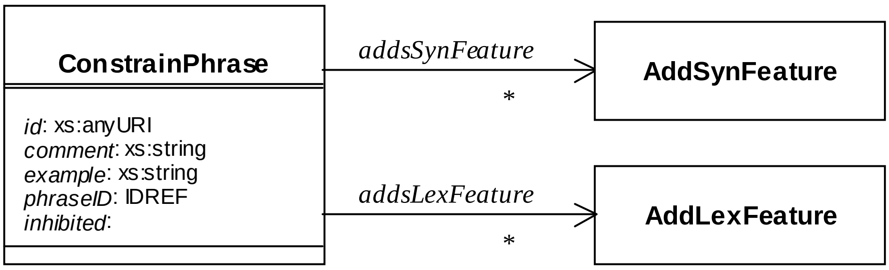
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <AddSlot>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <AddSlot>	xs:string	IMPLIED		
example	one or more example of the <AddSlot>	xs:string	IMPLIED		
slotID	the number of the new <Slot>	NMTOKEN	REQUIRED		

6.3.5.10 CONSTRAINPHRASE

This object constrains an existing <Phrase> by specifying new syntactic features

Diagram



Attributes

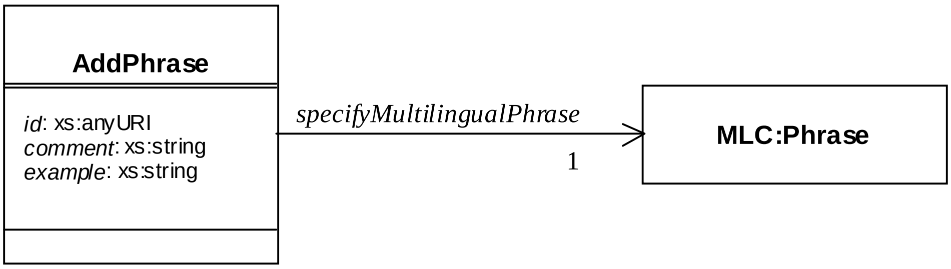
Name	Description	Type	Use	Default	Fixed

id	a unique identifier of the <ConstrainPhrase>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <ConstrainPhrase>	xs:string	IMPLIED		
example	one or more example of the <ConstrainPhrase>	xs:string	IMPLIED		
phraseID	the ID of the constrained <phrase>	IDREF	REQUIRED		
inhibited	whether its presence in the position is blocked or not	YES, NO		NO	

6.3.5.11 AddPhrase

This object adds a new <Phrase> filling a <Slot>.

Diagram



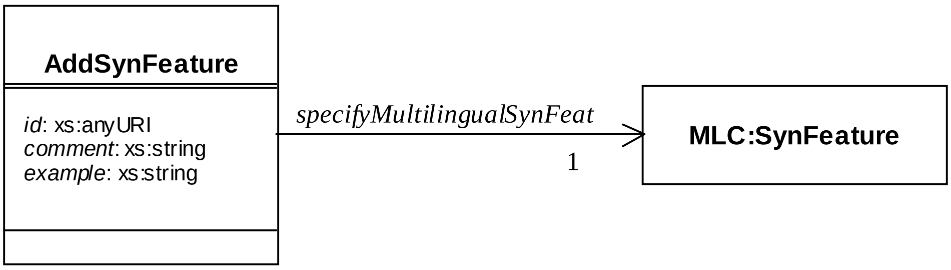
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <AddPhrase>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <AddPhrase>	xs:string	IMPLIED		
example	one or more example of the <AddPhrase>	xs:string	IMPLIED		

6.3.5.12 AddSynFeature

This object adds a new <SynFeature>.

Diagram



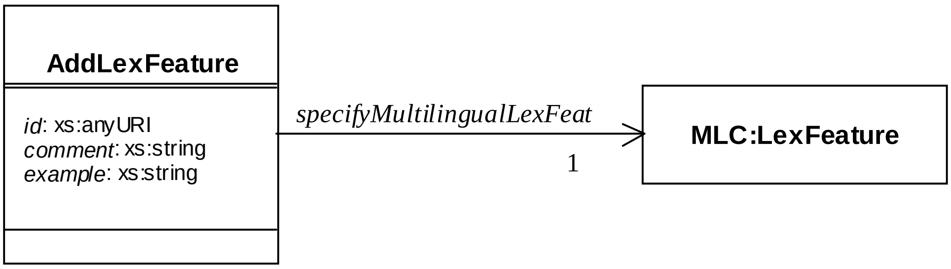
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <AddSynFeature>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <AddSynFeature>	xs:string	IMPLIED		
example	one or more example of the <AddSynFeature>	xs:string	IMPLIED		

6.3.5.13 AddLexFeature

This object adds a new <LexFeature>.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <AddLexFeature>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <AddLexFeature>	xs:string	IMPLIED		
example	one or more example of the <AddLexFeature>	xs:string	IMPLIED		

6.3.5.14 SlotMultCorresp

Relates a slot in the source <SynU> to one slot in the target <SynU>.

Attributes

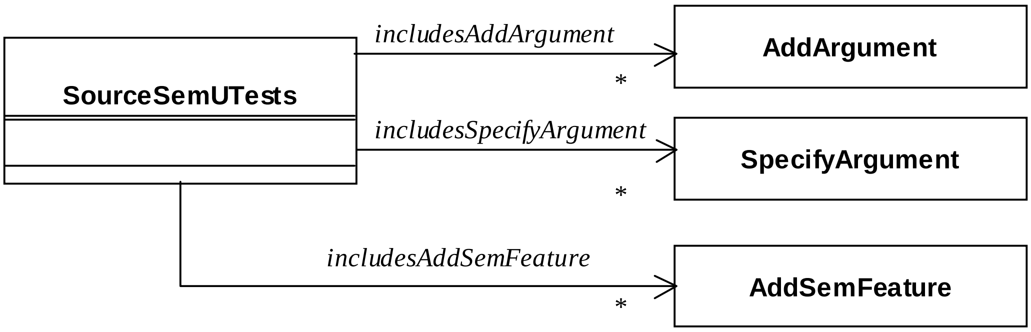
Name	Description	Type	Use	Default	Fixed
sourceSlot	the number of the source <Slot>	NMTOKEN	REQUIRED		
sourceSlotLocation	whether the linked slot is external to the <Self> (i.e. belongs to the <SynU> construction) or internal to the <Self>.	EXTERNAL, INTERNAL		EXTERNAL	
targetSlot	the number of the	NMTOKEN	REQUIRED		

	target <Slot>		
targetSlotLocation	whether the linked slot is external to the <Self> (i.e. belongs to the <SynU> construction) or internal to the <Self>.	EXTERNAL, INTERNAL	EXTERNAL

6.3.5.15 SOURCESEMUTESTS

This lexical object contains the semantic tests and constrains on the source <SemU>

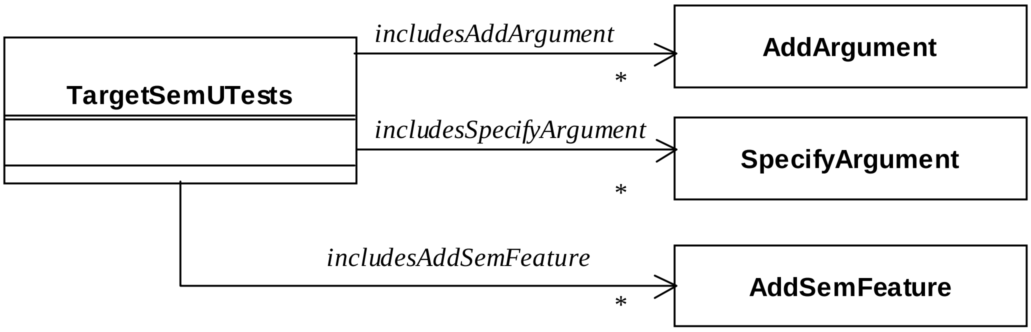
Diagram



6.3.5.16 TARGETSEMUACTIONS

This lexical object contains the syntactic actions on the target <SynU>

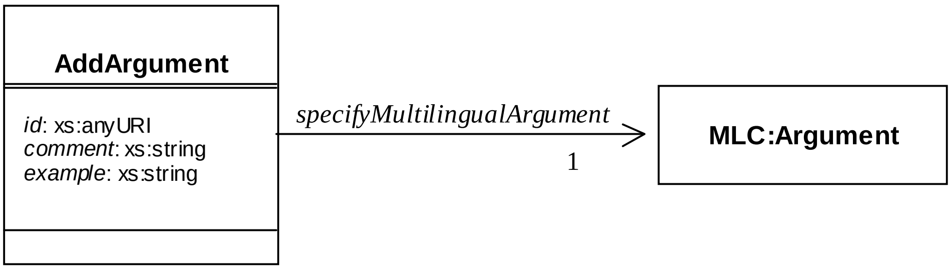
Diagram



6.3.5.17 ADDARGUMENT

This object adds a new <Argument> to an existing <SemanticFrame>.

Diagram



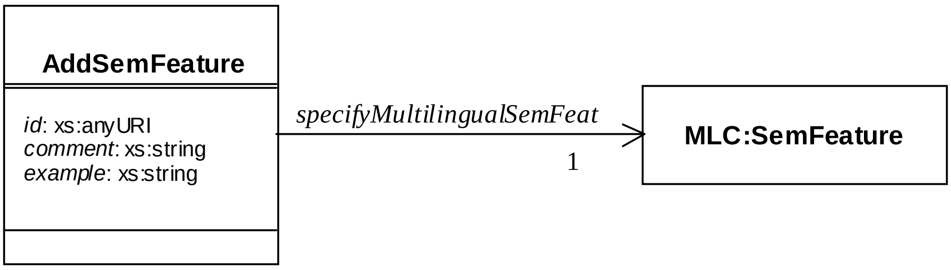
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <AddArgument>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <AddArgument>	xs:string	IMPLIED		
example	one or more example of the <AddArgument>	xs:string	IMPLIED		

6.3.5.18 ADDSEMFEATURE

This object adds a new <SemFeature>.

Diagram



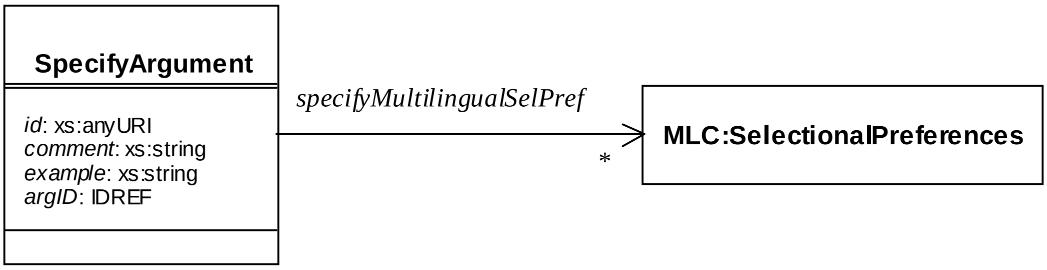
Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <AddSemFeature>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <AddSemFeature>	xs:string	IMPLIED		
example	one or more example of the <AddSemFeature>	xs:string	IMPLIED		

6.3.5.19 SPECIFYARGUMENT

Semantically specifies an existing semantic argument.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <AddSemFeature>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <AddSemFeature>	xs:string	IMPLIED		
example	one or more example of the <AddSemFeature>	xs:string	IMPLIED		
argID	the ID of the <Argument> to be specified	IDREF	REQUIRED		

6.3.5.20 ArgMultCorresp

Relates an argument in the source <SemU> to one argument in the target <SemU>.

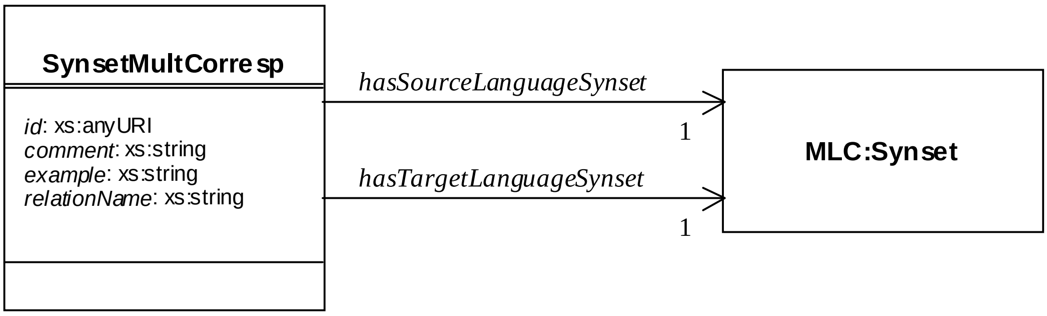
Attributes

Name	Description	Type	Use	Default	Fixed
sourceArg	the number of the source <Argument>	NMTOKEN	REQUIRED		
targetArg	the number of the target <Argument>	NMTOKEN	REQUIRED		

6.3.5.21 SynsetMultCorresp

This object specifies a link between synsets belonging to two different languages.

Diagram



Attributes

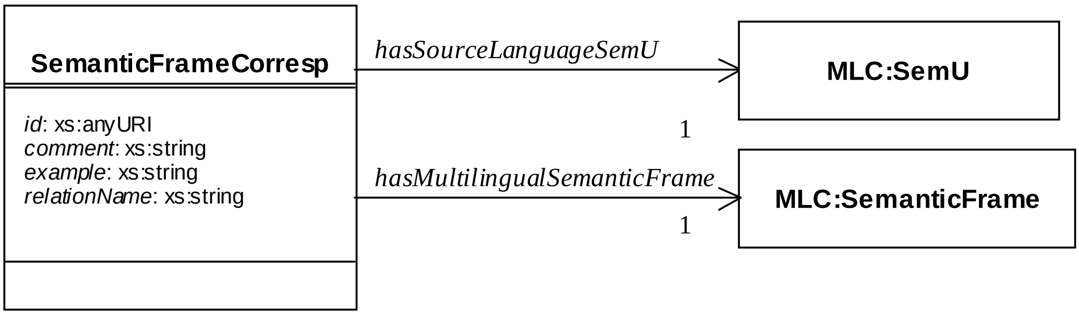
Name	Description	Type	Use	Default	Fixed

id	a unique identifier of the <SynsetMultCorresp>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SynsetMultCorresp>	xs:string	IMPLIED		
example	one or more example of the <SynsetMultCorresp>	xs:string	IMPLIED		
relationName	the name of the <SynsetMultCorresp>	xs:string	REQUIRED		

6.3.5.22 SEMANTICFRAMECORRESP

This object specifies a link between a source <SemU> and an interlingua <SemanticFrame>.

Diagram



Attributes

Name	Description	Type	Use	Default	Fixed
id	a unique identifier of the <SemanticFrameCorresp>	xs:anyURI	REQUIRED		
comment	a comment or short description of the <SemanticFrameCorresp>	xs:string	IMPLIED		
example	one or more example of the <SemanticFrameCorresp>	xs:string	IMPLIED		

6.4 Formalisation of MILE

6.4.1 Overview

The eventual vision for computational lexicons is to enable universal access to sophisticated linguistic information. Furthermore, for language processing applications (especially multi-lingual applications), it is desirable to provide means for *inferencing* over lexical information to determine its relevance for interpretation in a specific context.

The Resource Definition Framework (RDF) and the Ontology Web Language (OWL) have recently been developed by the World Wide Web Consortium (W3C). These two standards build upon the XML web infrastructure to enable the creation of a *Semantic Web*, wherein web objects can be classified according to their properties, and the semantics of their relations (links) to other web objects can be precisely defined. This in turn will enable powerful inferencing capabilities that can adapt processes to particular contexts.

The MILE lexical entry is an ideal structure for rendering in RDF. It consists of a hierarchy of lexical objects that are built up by combining atomic data categories via clearly defined relations. If mono- and multi-lingual lexical information can be eventually incorporated into the Semantic Web via its representation in RDF and OWL, it will provide an invaluable resource for language processing applications.

6.4.2 Proof of Concept

As a proof of concept, we have created an RDF schema for the ISLE/MILE lexical entry and instantiated one entry in several alternative forms to explore its potential as a representation for lexical data that can be integrated into the Semantic Web. The following describes the various components.

6.4.3 RDF schema for ISLE lexical entries

An RDF schema defines classes of objects and their relations to other objects. It does not in itself comprise an instance of these objects, but simply specifies the properties and constraints applicable to objects that conform to it.

A draft RDF schema for ISLE lexical entries is included in Appendix A. The classes and relations (properties) defined in the schema correspond to the ER diagrams (cf. 6.3). For example, the schema indicates that there is a class of objects called **Entry**; a property declaration indicates that the relation *hasSynU* holds between **Entry** objects and **SynU** objects. Note that classes can be defined to be subclasses of other classes, in which case properties associated with the parent class are inherited. In the ISLE schema, for example, the objects **Self** and **SlotRealization** are defined to be sub-classes of **PhraseElement**, and the *hasPhrase* property holds between any object of type **PhraseElement** (including its sub-classes) and objects of type **Phrase**.

The ISLE RDF schema and entries have been validated using the ICS-FORTH Validating RDF Parser (VRP v2.1), which analyzes the syntax of a given RDF/ XML file according to the *RDF Model and Syntax Specification* (http://www.w3.org/TR/rdf-syntax-grammar/), and checks whether the statements contained in both RDF schemas and resource descriptions satisfy the semantic constraints derived by the RDF Schema Specification (http://www.w3.org/TR/rdf-schema/).

6.4.4 ISLE Lexical Entries and the Data Category Registry

Appendix B contains three versions of the **SynU** description for "eat", instantiated as RDF objects. The first is a "full" version in which all of the information is specified, including atomic values (strings) at the leaves of the tree structure. The second two versions, rather than specifying all information explicitly, rely on the existence of a *Lexical Data Category Registry* (LDCR) in which pre-defined lexical objects are instantiated and may be included in the entry by a direct reference.

The potential to develop a Lexical Data Category Registry in which lexical objects are instantiated in RDF is one of the most important for the creation of multi-lingual, reusable lexicons. It allows for the following:

1. specification of a universally accessible, standard set of morphological, syntactic, and semantic information that can serve as a reference for lexicons creators;
2. a fully modular specification of lexical entities that enables use of all or parts of the lexical information in the repository as desired or appropriate, to build more complex lexical information modules;
3. means to reuse lexical specifications in entries sharing common properties, thereby eliminating redundancy as well as providing direct means to identify lexical entries or sub-entries with shared properties;
4. a universally accessible set of lexical information categories that may be used in applications or resources other than lexicons.

Note that the existence of a repository of lexical objects, instantiated and specified at different levels of complexity, does not imply that these objects must be used by lexicon creators. Rather, it provides a set of "off the shelf" lexical objects which either may be used as is, or which provide a departure point for the definition of new or modified categories.

The examples in Appendix B provide a small example of how a repository of RDF-instantiated lexical objects can be used. Sample repositories of lexical objects at three different levels of granularity, corresponding to the examples in Appendix B, are given in Appendix C:

1. a repository of *enumerated classes* for lexical objects at the lowest level of granularity; this comprises a definition of sets of possible values for various lexical objects. Any object of this type must be instantiated with one of the listed values.
2. a repository of *phrase classes* which instantiate common phrase types, e.g., NP, VP, etc.
3. a repository of *constructions* containing instantiations of common syntactic constructions (e.g., for verbs which are both transitive and intransitive, as shown in the example);
4. a template that lexicon creators can use to create their own data categories at any level of granularity.

The example entries demonstrate three different possibilities for the use of information in the repositories:

1. Entry 1 uses only the enumerated classes in the LDCR for **SynFeatureName** and **SynFeatureValue**. Note that in this case, the LDCR only provides a closed list of possible values, from which the assigned value in the entry must be chosen.
2. Entry 2 refers to instances of *phrase objects* in the LDCR rather than including them in the entry; this enables referring to a complex phrase (**Vauxhave** in the example) rather than including it directly in the entry, and provides the potential to reuse the same instance by reference in the same or other entries (this is done with **NP** in the example).

3. Entry 3 takes advantage of *construction instances* in the LDCR, thus eliminating the full specification in the entry and, again, allowing for reuse in other entries.

6.4.5 Purpose of the formalization

This is a first draft intended to exemplify how RDF may be used to instantiate lexical objects at various levels of granularity, which can be used and reused to create lexical entries within a single lexicon as well as across lexicons. By relying on the developing standardized technologies underlying the Semantic Web, we ensure universal accessibility and commonality. Ultimately, lexical objects defined in this way can be used not only for lexicons, but also in language processing and other applications.

This example serves primarily as a proof of concept that may be refined and modified as we consider in more depth the exact RDF representation that would best serve the needs of lexicon creation. However, the potential of exploiting the developments in the Semantic Web world for lexicon development should be clear. The following is a (partial) list of the aspects which need refinement and/or modification:

1. limit data range values for numbers, etc. (XML Schema DataTypes)
2. check on means to avoid creating classes to group bits of information (may be able to do this with an RDF Description and ID attribute, as long as the properties can be associated with any resource—but this limits validatability)
3. look into OWL mechanisms for more detailed specification of enumerations, restrictions on the uniqueness of properties, etc. (OWL currently not supported for validation so left out of the main schema)
4. create SemU representation with links into the SUMO and/or other generally available ontology.

7 The lexical tool

The aim of ISLE is to develop, disseminate and promote de facto HLT standards and guidelines for language resources, but also to develop a prototype tool to assist the development of multilingual lexical resources in MILE format.

The aim of this prototype tool is to

- exemplify and disseminate the MILE entry using an actual model with already existing monolingual data,
- assist the development of multilingual lexical resources following MILE schema,
- make extensive use of already existing PAROLE and SIMPLE lexicons,
- eventually test the goodness of the guidelines by using them in a real scenario.

Three aspects crucially determined the definition of the lexicographic station development platform we are describing here:

1. MILE is built as an additional layer on top of monolingual descriptions. In most cases, these monolingual layers already exist and need to be reused.
2. MILE is a general schema liable to be customized according to in-house needs in real scenarios.
3. The definition of the prototype tool an the definition of MILE itself were parallel tasks within the ISLE project. This means we had not a final model while developing the tool.

This situation led us to define a lexicographic station development platform that guarantees the portability of the final prototype. The lexicographical station development platform has been designed as a Tool builder which parses any DTD describing an Entity Relationship model in order to automatically (i) map the DTD into a relational dB, (ii) build up a user-friendly interface able to cover the most common requirements of a lexicographic station, and (iii) exemplify, test and validate the goodness of the MILE model in a real scenario, that is, reusing already existing monolingual resources such as PAROLE and SIMPLE lexicons.
This strategy will guarantee the portability of the final prototype to the final specifications, the portability of already existing resources (i.e. reusability of already existing monolingual resources) and, finally, the portability to specific applications thus allowing for customization.

This project lays on the idea that the information contained in a DTD which describes a conceptual model expressed in terms of Entity-Relationship Model can be used to automatically build up a relational dB. This ' DTD approach' allows easy customization. The user no longer has to accommodate to the structure and insights of the lexicographic tool but rather the tool accommodates to the requirements and idiosyncrasies of the user needs.

Since the resulting prototype tool is seen as an exemplification, it has been designed as a self-explaining tool and, therefore, the user can consult the mappings between the dB and the DTD and is provided with a set of browsing facilities that allow to understand the model behind.

7.1 General architecture

The lexicographical station development platform is a prototype tool generator that reads and parses a sgml/xml DTD and generates a relational data base that can be managed with a core web dB interface.

The lexicographical station development platform guarantees that already existing monolingual resources expressed in sgml/xml can be easily reused and ported by and to MILE.

Basically, the lexicographic station development platform includes a generation module, a customization module and a core interface module as exemplified below:

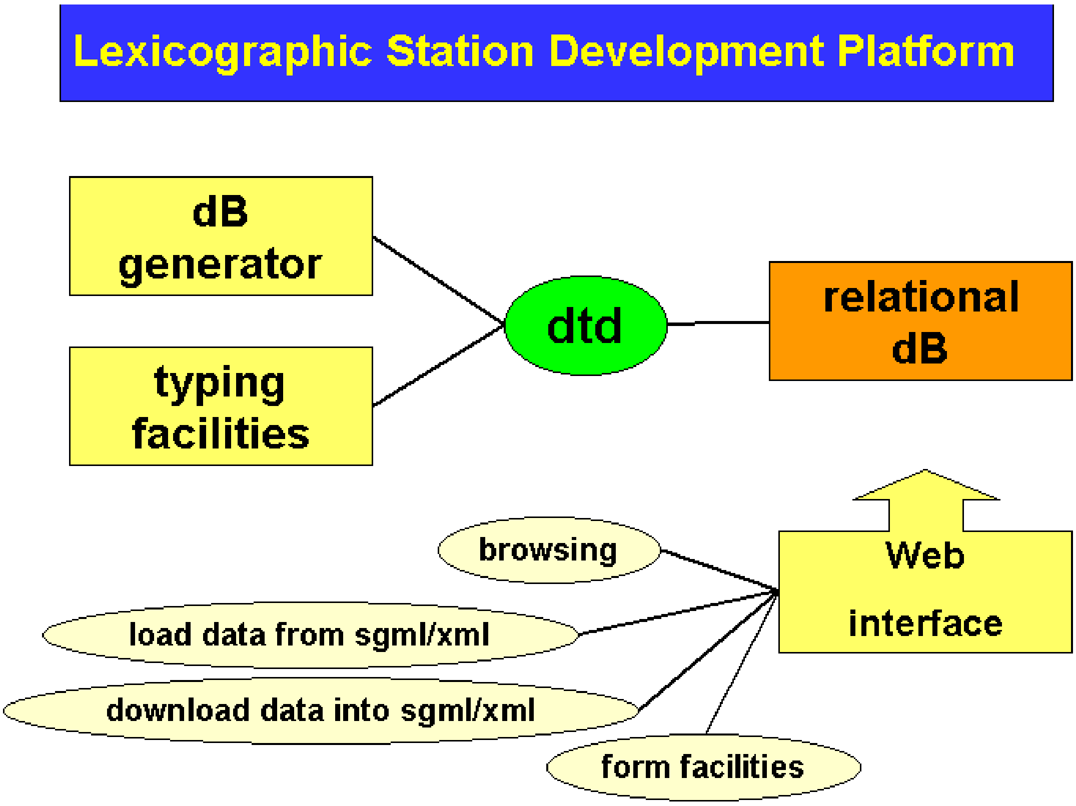


Figure 31

The **generation module** automatically generates a relational dB out of a DTD. The project benefits from the fact that a conceptual model expressed in terms of Entity-Relationship model can be easily mapped into a relational dB.

The **customisation module** allows to modify certain aspects of the dB at the time that allows overcoming some of the well known shortcomings of sgml DTDs such as typed references and type declaration.

The **core interface module** consists of a series of scripts that allow managing the dB with a friendly interface. Although user requirements differ from site to site according to in-house needs, the tool comes equipped with a set of basic functionalities. Our experience in past lexicographic projects led us to define an accurate list of requirements which include (i) query and browsing facilities, (ii) import, export and migration of data, (iii) easy encoding of new data, (iv) test and validation of both the data and the model, (v) customization facilities, and (vi) lexicographic tools such as type definition, class extraction and statistical facilities. As in the case of the generation module, this core interface module operates upon the model expressed in the DTD in order to make the necessary calculations to access, manipulate and display data from relevant tables.

ISLE defines the multilingual layer as an additional layer on top of the monolingual ones. Thus, whereas monolingual layers collect morphological, syntactic and semantic information needed to describe monolingual lexicons, the multilingual layer defines correspondence objects that describe relations between monolingual representations. This approach guarantees the independence of monolingual descriptions at the time that allows the maximum degree of flexibility. This structure can be represented as follows:

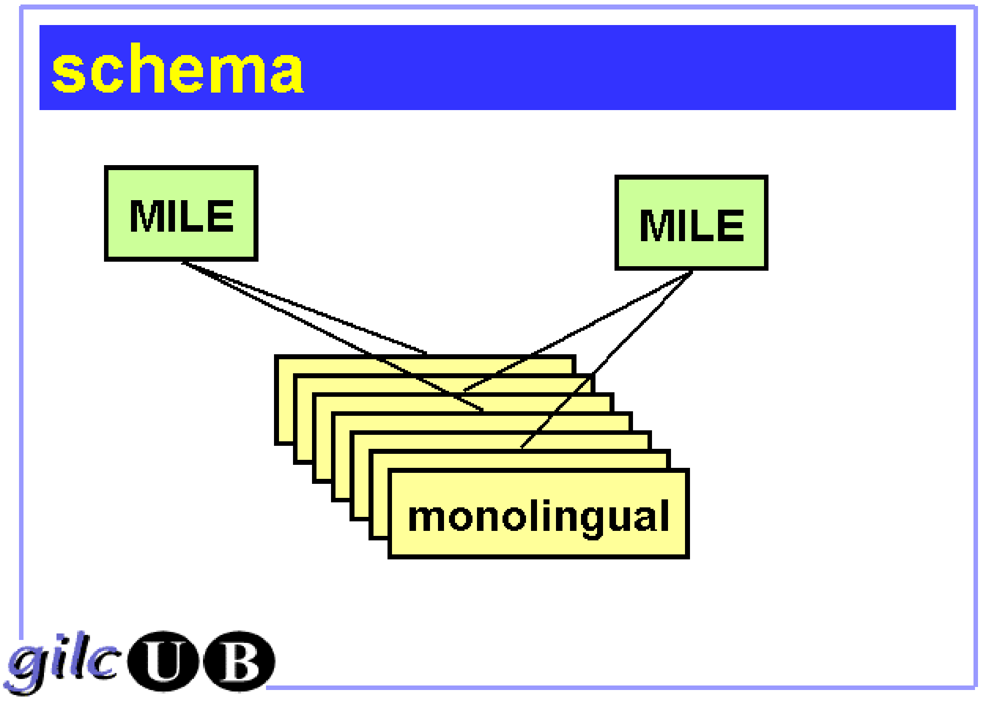


Figure 32

As we can see from figure above, the dB generator needs to generate at least two monolingual databases and one bilingual database and the web interface needs to address three different databases. The overall architecture of the system can be represented as follows:

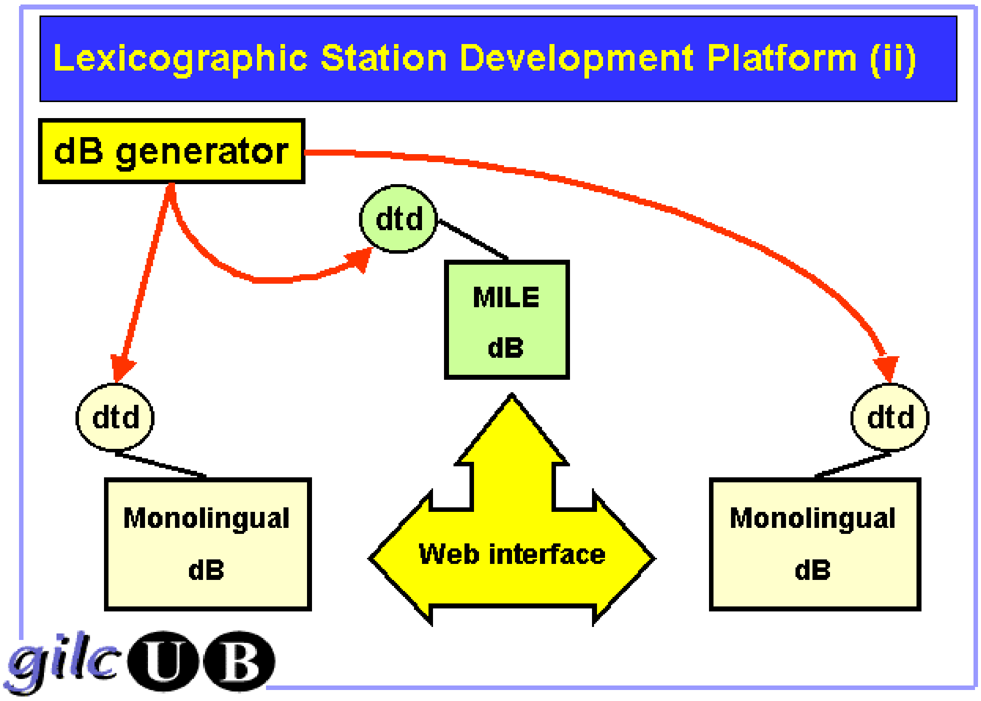

Lexicographic Station Development Platform (ii)
dB generator
dtd
MILE
dB
dtd
dtd
Monolingual
dB
Web interface
Monolingual
dB
gilc UB

Figure 33

8 Conclusions: Looking at the Future while Preserving the Past

Within ISLE we have worked looking at the future of lexicons, but building on past initiatives, i.e. with the aim of maximising the reuse of existing lexicons within a new framework.

The reuse of existing lexicons is achieved through the design of a lexical schema, the MILE, which (i) takes into consideration the basic notions employed in major available lexicons, (ii) is flexible enough to allow mapping from various lexical models into it, and (iii) allows the creation of user defined lexical objects if needed.

What is however of major importance is the attention paid – in designing the MILE - on recent developments particularly in the Semantic Web community, its standards and requirements. Lexicons will undoubtedly form an essential component and a building block of great impact to make the vision of a European pervasive Information Infrastructure and of the Semantic Web a reality. Language - and lexicons - are the gateway to knowledge. Lexicons - especially within a multilingual dimension - are at the base of bridging the knowledge gap in a multilingual society such as Europe: only through them can we tackle the twofold challenge of digital content availability and multilinguality. Only the development of integrated frameworks that include the activities of many will allow us to achieve a real break-through. Semantic Web developers will need repositories of words and terms - and knowledge about their relations within language use and ontological classification. The cost of adding this structured and *machine-understandable lexical information* can be one of the factors that delays its full deployment. But linguists alone will not be able to solve this. Like with the web (where many contribute), we have to get many people involved to make steps forward. A ***radical shift in the lexical paradigm - whereby many participants add linguistic content descriptions in an open distributed lexical framework*** - is required and proposed to make the Web usable.

8.1 MILE and the Semantic Web

The MILE Lexical Classes define the lexical objects to be used in building MILE conformant lexical entries, according to the MILE entry schema. Lexical objects include semantic and syntactic features, semantic relations, syntactic constructions, predicate and arguments, etc. The specifications of the Lexical Classes act as class definitions in an object-oriented language. Lexical Classes are organized in a hierarchy and defined using RDF schema (Brickley and Guha 2000), to formalize their properties and make their "semantics" explicit.
The MILE Data Categories represent instances of MILE Lexical Classes. They *form a first repository of recommended lexical objects*, selected for their lexicographic relevance or because they represent *de facto* standards in the NLP community. Users will be able to define new instances of lexical objects for their lexicon or language specific needs. This way, both at the monolingual and at the multilingual level (but with particular emphasis on the latter), ISLE intends to start up the incremental definition of a more object oriented architecture for lexicon design. Developers will be able to develop their own lexicon project either by selecting some of the MILE Data Categories or by defining new MILE conformant objects, which in turn might then enrich the common core if they reach a certain amount of consensus in the field. Data Categories will be identified by a URI and will act as common resources for lexical representation, to be in turn described by RDF

metadata. This way ISLE intends to foster the vision of open and distributed lexicons, with elements possibly residing in different sites on the Web. RDF descriptions and common definition will grant lexical content interoperability, enhancing the re-use and sharing of lexical resources and components.

8.2 Towards a new paradigm for Lexical Resources

The new foreseen and proposed paradigm of an **Open Lexical Infrastructure** requires new approaches at various levels, and lexicon creation, updating and maintenance will leave the comparatively few offices of linguists, to involve broad groups of experts, classrooms, but also the general public. The effort of making available millions of 'words' for dozens of languages is something that no individual or small group is able to afford. It has already been proved by a number of projects that lexicon building and maintenance can be achieved in a cooperative way. We think that it is now time to broaden and open the concept of cooperative effort to a much larger set of communities.

If we look at the *past*, in the last decade many activities, at European level and world-wide, have contributed to substantially advance knowledge and capability of how to represent/create/maintain/acquire/access/tune large lexical repositories. These repositories are rich in linguistic knowledge (and often in world knowledge), and many are based on best practices and standards that have been consensually agreed on or have been submitted to the international community as de facto standards. Core - or even large - lexical repositories have been and are being built for many European (and non EU) languages. Most came into existence in European projects, and continued in National Projects, thus creating the necessary platform for a future European lexical infrastructure. European researchers have played an outstanding role in these initiatives. An increasing software library allows integrated access to these resources and the creation of new material.

Looking at the *future,* a further step and *radical change of perspective* is now needed in order *to facilitate the integration of the linguistic information resulting from all these initiatives, to bridge the differences between various perspectives on language structure and linguistic content, to put an infrastructure into place for content description and content interoperability at European level and beyond, and to make lexical resources usable within the emerging Semantic Web scenario.* This objective can only be achieved when working in the direction of an integrated **Open and Distributed Lexical Infrastructure**, based on open content interoperability standards, where not only the linguistic experts can participate, but which includes designers, developers and users of content encoding practices, and also many members of the society.

We have designed *MILE* in such a way that it can serve *as a basic platform* for this Lexical Open and Distributed Infrastructure.

The approach foreseen to achieve the objectives requires, among others, the coverage of a range of aspects pertaining to linguistic modeling, and a number of organisational aspects, such as:

- The design of an abstract model of lexicon architecture (based on MILE) that offers the structural bandwidth necessary to include the various contributions, to allow building/maintaining/accessing/tuning… such complex, shared and distributed lexical repositories. This will, amongst others, entail the design and implementation of new models for linking monolingual lexicons and creating multilingual correspondences. Overall, there will be a tendency towards increasingly complex lexicons that reflect shared conceptual models. On the other side structural flexibility must be ensured, so as to allow easy and varied import and export of various lexicon types (from very complex to very simple). Many uses/users may require, in fact, simple lexicons with simple information types. We foresee *an increasing number of well-defined linguistic data categories stored in open and standardized repositories which will be used by users to define their own structures within an open lexical framework.* It is this re-usage of linguistic objects which will link new contents to the already existing lexical objects.

- The standardization effort will involve the extension and integration of existing and emerging open lexical and terminological standards and best practices such as EAGLES, ISLE, TEI, OLIF, Martif (ISO 12200) and Data Categories (ISO 12620). Initiatives towards the creation of lexical metadata such as IMDI, Dublin Core and OLAC will be taken into account. MILE may form the core part of the European contribution to the ISO-revision process and form the basis for future lexical standardization in the ISO/TC 37/SC 4 Committee.
- The *fostering of language resources integration and interoperability through links to these standards*. In the model of open data categories as primitive and constructed forms, consensual core basic lexical notions are implemented as basic shared lexical 'objects'. Users will be encouraged to use these lexical objects for the description, creation, management, and delivery of their resources, therefore creating ***semantic interoperability***. New objects can be created and linked up to the core set. This will ensure a *flexible model while working with a core set of lexical data categories*. It will guarantee freedom for the user to add or change objects if that is deemed necessary, provide an evaluation protocol for the core standard lexical data categories, and require verification methods for the integration of new objects.
- With MILE we put the basis for the realization of a common platform for *interoperability between different fields of linguistic activity* - such as lexicology, lexicography, terminology - *and Semantic Web development*. The platform will provide a flexible common environment not only for linguists, terminologists and ontologists, but also for content providers and content management software vendors, for development and communication. This will enable users to share lexicons and collaborate on parts of it. The lexicons may be distributed, i.e. different building blocks may reside at different locations on the web and are linked by URLs. This is strictly related to the Semantic Web standards (e.g. RDF metadata to describe lexicon data categories). Overall, lexicons will perform the bridging function between documents and conceptual categorization. The common conceptual model within the envisaged architecture will ensure content interoperability between texts, lexicons and ontologies.

Semantic content processing lies at the heart of the Semantic Web enterprise, and requires to squarely address the complexity of natural language. Existing experience in language resource development proves that such a challenge can be tackled only by pursuing a truly interdisciplinary approach, and by establishing a highly advanced environment for the representation and acquisition of lexical information, open to the reuse and interchange of lexical data.

Coming from the experience gathered in developing advanced lexicon models such as the SIMPLE one, and along the lines pursued by the ISLE standardization process, a ***new generation of lexical resources*** can be envisaged. These will crucially provide the semantic information to necessary allow for effective content processing. On the other hand, they *will in turn benefit from the Semantic Web itself*. Thus, it is possible to state the existence of a ***bi-directional relation between the Semantic Web enterprise and computational lexicon design and construction***. In fact, *the Semantic Web is going to crucially determine the shape of the language resources of the future.* Semantic Web emerging standards, such as ontologies, RDF, etc., allow for a new approach to language resource development and maintenance, which is consistent with the vision of an open space of sharable knowledge available on the Web for processing

References

Atkins, B. T. S. (1993) "Theoretical Lexicography and its Relation to Dictionary-making". In *Dictionaries: the Journal of the Dictionary Society of North* America, (guest editor) W. Frawley, DSNA, Cleveland Ohio. pp. 4-43.

Atkins, B. T. S. (1996) "Bilingual Dictionaries: Past, Present and Future". In: *Euralex' 96 Proceedings*, (eds.) Gellerstam, M. , J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström and C. R. Papmehl. Gothenburg: Gothenburg University, Department of Swedish. pp. 515-590.

Atkins, B. T. S. & K. Varantola (1997) "Monitoring Dictionary Use". In *International Journal of Lexicography*, 10:1, 1-45, and reprinted (1998) in *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*, ed. Atkins B. T. S. Tübingen : Niemeyer.

Atkins, B. T. S. & K. Varantola (1998) "Language Learners Using Dictionaries: The Final Report of the EURALEX- and AILA-sponsored Research Project into Dictionary Use". In *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*, ed. Atkins B. T. S. Tübingen : Niemeyer.

Atkins, B. T. S. & Nuria Bel, Francesca Bertagna, Pierrette Bouillon, Nicoletta Calzolari, Christiane Fellbaum, Ralph Grishman, Alessandro Lenci, Catherine MacLeod, Martha Palmer, Gregor Thurmair, Marta Villegas, Antonio Zampolli (2000) "From Resources to Applications. Designing the Multilingual ISLE Lexical Entry", LREC2002, Las Palmas.

Bel N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. *LREC Proceedings*, Athens.

Burnard, L., Baker, P., McEnery, A. & Wilson, A. (1997). *An analytic framework for the validation of language corpora*. Report of the ELRA Corpus Validation Group.

Calzolari, N. (1998). An Overview of Written Language Resources in Europe: a few Reflections, Facts, and a Vision, in A. Rubio, N. Gallardo, R. Castro, A. Tejada (eds.), *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, pp.217-224.

Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli A. (2002). *Towards Best Practice for Multiword Expressions in Computational Lexicons*. In Proceedings of LREC 2002, Las Palmas, Canary Islands, Spain.

Calzolari, N., Grishman, R., Palmer, M. (eds.) (2001a). *Survey of major approaches towards Bilingual/Multilingual Lexicons.* ISLE Deliverable D2.1-D3.1, Pisa.

Calzolari, N., Lenci, A., Zampolli, A., Bel, N., Villegas, M., Thurmair, G. (2001b). *The ISLE in the Ocean. Transatlantic Standards for Multilingual Lexicons (with an eye to Machine Translation).* In Proceedings of Machine Translation Summit VIII, Santiago de Compostela, Spain.

Calzolari, N., Mc Naught, J., Zampolli, A. (1996). *EAGLES Final Report: EAGLES Editors' Introduction*. EAG-EB-EI, Pisa.

Cobuild: English Dictionary. HarperCollins, London.

Collins Gem English-French Dictionary. HarperCollins, London.

The Concise Oxford Dictionary Oxford University Press: Oxford.

Cruse, A. (1986). *Lexical Semantics*. CUP, Cambridge UK.

Dorr, B.J., Jordan, P. W., and Benoit, J. W, "A Survey of Current Research in Machine Translation," Advances in Computers, Vol 49, M. Zelkowitz (Ed), Academic Press, London, pp. 1--68, 1999.

EAGLES (1996). *Evaluation of Natural Language Processing Systems.* Final Report, Center for Sprogteknologi, Copenhagen. Also available at http://issco-www.unige.ch/projects/ewg96/ewg96.html.

Fellbaum, C. (ed.) (1998). *WordNet : an Electonic Lexical Database*. MIT, Cambridge MA.

Fillmore, Charles J. & B. T. S. Atkins (1998) "FrameNet and Lexicographic Relevance". In *Proceedings of the First International Conference On Language Resources And Evaluation*, Granada, Spain, 28-30 May 1998.

Fillmore, Charles J. & B. T. S. Atkins (2000) "Describing polysemy : the case of crawl". In *Polysemy: Linguistic and Computational Approaches*, (eds) Yael Ravin and Claudia Leacock. Oxford: Oxford University Press.

Fillmore, Charles J., B. T. S. Atkins & C. R. Johnson (forthcoming) "Lexicographic Relevance: Essential Facts about Headwords". In *International Journal of Lexicography*, Guest Issue on FrameNet.

Fontenelle, Thierry (1997) *Turning a bilingual dictionary into a lexical-semantic database*, Lexicographica Series Maior 79, Max Niemeyer Verlag, Tübingen, Germany

GENELEX Consortium, (1994). *Report on the Semantic Layer*, Project EUREKA GENELEX, Version 2.1.

Gibbon, D., Moore R., Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, New York.

Hanks, P., (2000). Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance, in: U. Heid et al. (eds.) *Proceedings of Ninth Euralex International Congress*, IMS Stuttgart University, Stuttgart, Germany, I:3-13.

Heid, U., McNaught, J. (1991). *EUROTRA-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications*. Final report.

Leech, G., Wilson, A. (1996). *Recommendations for the morphosyntactic annotation of corpora*, Eag-tcwg-mac/r, Lancaster.

Lenci A., Bel N., Busa F., Calzolari N., Gola E., Monachini M., Ogonowsky A., Peters I., Peters W., Ruimy N., Villegas M., Zampolli A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, XIII (4), 249--263.

Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., Zampolli, A. (1999). *Linguistic Specifications*. SIMPLE Deliverable D2.1. ILC and University of Pisa

Lieske, C.,McCormick, S., Thurmair, G. (2001). *The Open Lexicon Interchange Format (OLIF) Comes of Age*. In Proceedings of the Eighth Machine Translation Summit, Santiago de Compostela, Spain.

Monachini, M., Calzolari, N. (1996). *Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages*, Eag-clwg-morphsyn/r, ILC-CNR, Pisa.

Monachini, M., Calzolari, N. (1999). Standardization in the Lexicon, in H. van Halteren (ed.), *Syntactic Wordclass Tagging*, Kluwer, Dordrecht, pp. 149-173.

The *New Oxford Dictionary of English*, Oxford University Press, Oxford.

PAROLE-Corpus. Pisa: ILC-CNR.

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

Roland, & Jurafsky (1998) Verb-Sense and Verb-Subcategorization Probabilities. in Stevenson, S. and P. Merlo (eds.) *CUNY Sentence Processing Conference*, Benjamins

Roventini, A., Alonge A., Bertagna F., Calzolari N., Cancila J., Girardi, C., Magnini, B., Manrinelli R., Speranza, M., Zampolli, A. (2002) ItalWordNet: Building a Large Semantic Databaes for the Automatic Treatment of Italian. in *Rivista di Linguistica Computazionale* (in press).

Ruimy, N., Corazzari, O., Gola, E., Spanu, A., Calzolari, N., Zampolli, A. (1998). The European LE-PAROLE Project: The Italian Syntactic Lexicon, in *Proceedings of the First International Conference on Language resources and Evaluation*, Granada: 241-248.

Ruimy N., Monachini M., Distante R., Guazzini E., Molino S., Ulivieri M., Calzolari N., Zampolli A. (2002), "CLIPS, a Mult-level Italian Computational Lexicon: a Glimpse to Data, in

Proceedings of Third International Conference on Language Resources and Evaluation - LREC2002, vol.III, pp.792-799, Las Palmas de Gran Canaria, Spain.

Sanfilippo, A. *et al.* (1996). *EAGLES Preliminary Recommendations on Subcategorisation* EAG---CLWG---SYNLEX/P Version of Aug, 1996. 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

Sinclair, J. (1991) *Corpus, Concordance and Collocation* OUP, Oxford.

Thurmair, G. (2000). *OLIF Input Document*, June 2000. See http://www.olif.net/main.htm

Underwood, N. & Navarretta, C. (1997*). A Draft Manual for the Validation of Lexica*. Final ELRA Report, Copenhagen.

Villegas, M., Bel, N. (2002). From DTDs to relational dBs. An automatic generation of a lexicographical station out off ISLE guidelines. In Proceedings of LREC 2002, Las Palmas, Canary Islands, Spain.

Vossen, P. (ed.) (1999) *EuroWordNet General Document.* EuroWordnet (LE2-4003, LE4-8328), Final Document. www.illc.uva.nl/EuroWordNet/docs.html

Zampolli, A. (1997). The PAROLE project in the general context of the European actions for Language Resources, in R. Marcinkeviciene, N. Volz (eds.), *TELRI Proceedings of the Second European Seminar: Language Applications for a Multilingual Europe,* IDS/VDU, Manheim/Kaunas.

Zampolli, A. (1998). Introduction of the General Chairman, in A. Rubio, N. Gallardo, R. Castro, A. Tejada (eds.), *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada.

Appendix A : RDF Schema for ISLE/MILE Lexical Entries

```
<!--
       An RDF Schema for ISLE lexical entries

       v 0.6 2002/11/04
       Author: Nancy Ide

-->

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:owl ="http://www.w3.org/2002/07/owl#
         xmlns:mlc ="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#">

         <1-- ISLE/MILE lexical objects (classes) -->

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Entry">
<rdfs:label>Entry</rdfs:label>
<rdfs:comment>This class holds entries</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynU">
<rdfs:label>SynU</rdfs:label>
<rdfs:comment>This class holds syntactic information</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SemU">
<rdfs:label>SemU</rdfs:label>
<rdfs:comment>This class holds semantic information</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#MU">
<rdfs:label>MU</rdfs:label>
<rdfs:comment>This class holds morpho-syntactic information</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#FrameSet">
<rdfs:label>FrameSet</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SyntacticFrame">
<rdfs:label>description</rdfs:label>
<rdfs:comment>Holds subcategorization information</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Self">
<rdfs:label>Self</rdfs:label>
<rdfs:comment>Specifies   the   properties   of   the   head   of   the   syntactic
pattern</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Construction">
<rdfs:label>Construction</rdfs:label>
<rdfs:comment>Specifies the complementation pattern of Self</rdfs:comment>
<rdfs:subClassOf             rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#PhraseElement"/>
</rdfs:Class>

<rdfs:Class                     rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SlotRealization">
<rdfs:label>SlotRealization</rdfs:label>
<rdfs:comment>specifies ways the slot can be syntactically realized</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Phrase">
<rdfs:label>Phrase</rdfs:label>
<rdfs:comment>This class holds phrases</rdfs:comment>
<rdfs:subClassOf             rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#PhraseElement"/>
</rdfs:Class>
```

```
<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynFeature">
<rdfs:label>SynFeature</rdfs:label>
<rdfs:comment>This class holds feature-value pairs</rdfs:comment>
</rdfs:Class>

<!-- The following are not proper MILE lexical classes, but group information
     by allowing association of different kinds of info to a given node in the
     RDF realization. This needs to be looked into to see if there is a better
     way to do this
-->


<rdfs:Class                     rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelativeOrderConstraint">
<rdfs:label>RelativeOrderConstraint</rdfs:label>
<rdfs:comment>Groups together ordering constraint information</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#PhraseElement">
<rdfs:label>PhraseElements</rdfs:label>
<rdfs:comment>Things that have the slot property</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#RelatedSlots">
<rdfs:label>RelatedSlots</rdfs:label>
</rdfs:Class>


<!-- Properties (relations) between objects and between objects and atomic
     values
-->

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#hasSynu">
<rdfs:label>synu</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Entry"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynU"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#example">
<rdfs:label>points to examples</rdfs:label>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Resource"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>

<rdf:Property                   rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasSyntacticFrame">
<rdfs:label>description</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynU"/>
<rdfs:range                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SyntacticFrame"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#hasFrameSet">
<rdfs:label>frameSet</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynU"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#FrameSet"/>
</rdf:Property>

<rdf:Property                   rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#relatesFrames">
<rdfs:label>frameSet</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#FrameSet"/>
<rdfs:range                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SyntacticFrame"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#specifiedBy">
<rdfs:label>frameSet</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#FrameSet"/>
<rdfs:range                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#sourceSlot">
<rdfs:label>sourceSlot</rdfs:label>
<rdfs:domain                 rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Literal"/>
```

```
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#targetSlot">
<rdfs:label>targetSlot</rdfs:label>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Literal"/>
</rdf:Property>

<rdf:Property                    rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasSourceFrame">
<rdfs:label>hasSourceFrame</rdfs:label>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
<rdfs:range                   rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SyntacticFrame"/>
</rdf:Property>

<rdf:Property                    rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasTargetFrame">
<rdfs:label>hasTargetFrame</rdfs:label>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
<rdfs:range                   rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SyntacticFrame"/>
</rdf:Property>

<rdf:Property                    rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#selectSourcePhrase">
<rdfs:label>selectSourcePhrase</rdfs:label>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Phrase"/>
</rdf:Property>

<rdf:Property                    rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#selectTargetPhrase">
<rdfs:label>selectTargetPhrase</rdfs:label>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Phrase"/>
</rdf:Property>

<rdf:Property                    rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasTargetFrame">
<rdfs:label>hasTargetFrame</rdfs:label>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelatedSlots"/>
<rdfs:range                   rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SyntacticFrame"/>
</rdf:Property>

<rdf:Property                    rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#correspondsTo">
<rdfs:label>CorrespSynUSemU</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynU"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SemU"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#composedBy">
<rdfs:label>composition</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynU"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#MU"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#hasSelf">
<rdfs:label>hasSelf</rdfs:label>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SyntacticFrame"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Self"/>
</rdf:Property>

<rdf:Property                    rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasConstruction">
<rdfs:label>hasConstruction</rdfs:label>
<rdfs:comment>Used to encode MWEs</rdfs:comment>
<rdfs:domain                  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SyntacticFrame"/>
```

```
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#Construction"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#hasFrequency">
<rdfs:label>frequency</rdfs:label>
<rdfs:comment>specifies the frequency in the corpus</rdfs:comment>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Resource"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/> <!-- number -->
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#corpus">
<rdfs:label>corpus</rdfs:label>
<rdfs:comment>specifies the corpus upon which the frequency is based</rdfs:comment>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Resource"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#headedBy">
<rdfs:label>head of phrase</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Self"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Phrase"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasInternalConstruction">
<rdfs:label>construction</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Self"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#Construction"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#orderedBy">
<rdfs:label>order constraints</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#Construction"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#RelativeOrderConstraint"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#position">
<rdfs:label>construction</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#Construction"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#slot">
<rdfs:label>realization</rdfs:label>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#PhraseElement"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SlotRealization"/>
</rdf:Property>

<owl:FunctionalProperty rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasFunction">
<rdfs:label>function</rdfs:label>
<rdfs:comment>specifies the head</rdfs:comment>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SlotRealization"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-enumerated-
classes#FunctionType"/>
</owl:FunctionalProperty>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#filledBy">
<rdfs:label>phrase</rdfs:label>
<rdfs:comment>specifies </rdfs:comment>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SlotRealization"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Phrase"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#lexicalizedBy">
<rdfs:label>lexicalizeBy</rdfs:label>
<rdfs:comment>lexicalization of the phrase</rdfs:comment>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Phrase"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#MU"/>
```

```
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasSynFeature">
<rdfs:label>feature</rdfs:label>
<rdfs:comment>specifies the feature-value pairs</rdfs:comment>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#Phrase"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#SynFeature"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasSynFeatureName">
<rdfs:label>featureName</rdfs:label>
<rdfs:comment>specifies the feature name</rdfs:comment>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SynFeature"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-enumerated-
classes#SynFeatureName"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#hasSynFeatureValue">
<rdfs:label>featureValue</rdfs:label>
<rdfs:comment>specifies the feature value</rdfs:comment>
<rdfs:domain rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-schema-
v.6#SynFeature"/>
<rdfs:range rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-enumerated-
classes#SynFeatureValue"/>
</rdf:Property>

<!-- Some properties for specifying ordering constraints -- to be looked into
     for a possibly better way to handle
-->

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#beforeSlot">
<rdfs:label>beforeSlot</rdfs:label>
<rdfs:comment>the number of the slot occurring before in an ordering constraint
specification</rdfs:comment>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Resource"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#afterSlot">
<rdfs:label>afterSlot</rdfs:label>
<rdfs:comment>the number of the slot occurring after in an ordering constraint
specification</rdfs:comment>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Resource"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#beforePhrase">
<rdfs:label>beforePhrase</rdfs:label>
<rdfs:comment>phrase occurring before in an ordering constraint
specification</rdfs:comment>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Resource"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Phrase"/>
</rdf:Property>

<rdf:Property rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#afterPhrase">
<rdfs:label>afterPhrase</rdfs:label>
<rdfs:comment>phrase occurring after in an ordering constraint
specification</rdfs:comment>
<rdfs:domain rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Resource"/>
<rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Phrase"/>
</rdf:Property>

</rdf:RDF>
```

Appendix B: Sample Entries

ENTRY 1 : Full entry

Highlighted lines refer to objects whose values are constrained in LDCR definitions (Appendix C).

```
<?xml version="1.0"?>
<!--
     Sample ISLE lexical Entry for EAT (transitive), SynU only
     Abbreviated syntax version using no pre-defined objects
     2002/10/23 Author: Nancy Ide
-->
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:mlc="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#"
         xmlns="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#">

<Entry rdf:ID="eat1">

   <!-- The SynU for eat1 -->
   <hasSynu rdf:parseType="Resource">
      <SynU rdf:ID="eat1-SynU">
         <example>John ate the cake</example>
         <hasSyntacticFrame>
            <SyntacticFrame rdf:ID="eat1SynFrame">
               <hasSelf>
                  <Self rdf:ID="eat1Self">
                     <headedBy>
                        <Phrase rdf:ID="Vauxhave">
                           <hasSynFeature>
                              <SynFeature>
                                 <hasSynFeatureName rdf:value="aux"/>
                                 <hasSynFeatureValue rdf:value="have"/>
                              </SynFeature>
                           </hasSynFeature>
                        </Phrase>
                     </headedBy>
                  </Self>
               </hasSelf>
               <hasConstruction>
                  <Construction rdf:ID="eat1Const">
                     <slot>
                        <SlotRealization rdf:ID="NPsubj">
                            <hasFunction rdf:value="Subj"/>
                            <filledBy rdf:value="NP"/>
                        </SlotRealization>
                     </slot>
                     <slot>
                        <SlotRealization rdf:ID="NPobj">
                             <hasFunction rdf:value="Obj"/>
                             <filledBy rdf:value="NP"/>
                        </SlotRealization>
                     </slot>
                   </Construction>
                </hasConstruction>
                <hasFrequency rdf:value="8788" mlc:corpus="PAROLE"/>
            </SyntacticFrame>
         </hasSyntacticFrame>
       </SynU>
    </hasSynu>
 </Entry>
 </rdf:RDF>
```

ENTRY 2 : Using LDCR categories for PHRASE

The highlighted lines refer to pre-instantiated lexical objects. A portion of the LDCR for Phrases is given in Appendix C. The URL reference is to the actual web address where the object is instantiated.

```
<?xml version="1.0"?>
```

```
<!--
     Sample ISLE lexical Entry for EAT (transitive), SynU only
     Abbreviated syntax version using no pre-defined objects
     2002/10/23 Author: Nancy Ide
-->

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:mlc="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.4#"
         xmlns="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.4#">

<Entry rdf:ID="eat1">

   <!-- The SynU for eat1 -->

   <hasSynu rdf:parseType="Resource">
      <SynU rdf:ID="eat1-SynU">
         <example>John ate the cake</example>
         <hasSyntacticFrame>
            <SyntacticFrame rdf:ID="eat1SynFrame">
               <hasSelf>
                  <Self rdf:ID="eat1Self">
                     <headedBy
                        rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-
datcats/Phrases#Vauxhave"/>
                  </Self>
               </hasSelf>
               <hasConstruction>
                  <Construction rdf:ID="eat1Const">
                     <slot>
                       <SlotRealization rdf:ID="NPsubj">
                         <hasFunction rdf:value="Subj"/>
                         <filledBy  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-
datcats/Phrases#NP"/>
                       </SlotRealization>
                     </slot>
                     <slot>
                       <SlotRealization rdf:ID="NPobj">
                         <hasFunction rdf:value="Obj"/>
                         <filledBy  rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-
datcats/Phrases#NP"/>
                       </SlotRealization>
                     </slot>
                   </Construction>
                </hasConstruction>
                <hasFrequency rdf:value="8788" mlc:corpus="PAROLE"/>
            </SyntacticFrame>
         </hasSyntacticFrame>
       </SynU>
    </hasSynu>

</Entry>

 </rdf:RDF>
```

ENTRY 3 : Using LDCR categories for CONSTRUCTION

The highlighted lines refer to a pre-instantiated **Construction** object. A portion of the LDCR for Constructions is given in Appendix C. The URL reference is to the actual web address where the object is instantiated.

```
<?xml version="1.0"?>

<!--
     Sample ISLE lexical Entry for EAT (transitive)
     Abbreviated syntax version using pre-defined construction
     2002/10/23 Author: Nancy Ide
-->

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:mlc="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#"
         xmlns="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#">

<Entry rdf:ID="eat1">
```

```
   <!-- The SynU for eat1 -->

   <hasSynu rdf:parseType="Resource">
      <SynU rdf:ID="eat1-SynU">
         <example>John ate the cake</example>
         <hasSyntacticFrame>
            <SyntacticFrame rdf:ID="eat1SynFrame">
               <hasSelf>
                  <Self rdf:ID="eat1Self">
                     <headedBy
                      rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-
datcats/Phrases#Vauxhave"/>
                  </Self>
               </hasSelf>
               <hasConstruction
                rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-
datcats/Constructions#TransIntrans"/>
               <hasFrequency rdf:value="8788" mlc:corpus="PAROLE"/>
            </SyntacticFrame>
         </hasSyntacticFrame>
       </SynU>
    </hasSynu>

  </Entry>


  </rdf:RDF>
```

Appendix C: LDCR definitions

Sample LDCR entries specifying enumerated values for **SynFeatureName**, etc. The specification uses the Ontology Web Language (OWL) to list valid values for objects of the defined class.

```
<!--
       Enumerated classes for ISLE lexical entries
       v 0.1 2002/10/23
       Author: Nancy Ide
-->

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:owl ="http://www.w3.org/2002/07/owl#
         xmlns:isle ="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#">

<rdfs:Class                 rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-enumerated-
classes#FunctionType">
<owl:oneOf>
   <rdf:Seq>
      <rdf:li>Subj</rdf:li>
      <rdf:li>Obj</rdf:li>
      <rdf:li>Comp</rdf:li>
      <rdf:li>Arg</rdf:li>
      <rdf:li>Iobj</rdf:li>
   </rdf:Seq>
</owl:oneOf>
</rdfs:Class>

<rdfs:Class                 rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-enumerated-
classes#SynFeatureName">
<owl:oneOf>
   <rdf:Seq>
      <rdf:li>tense</rdf:li>
      <rdf:li>gender</rdf:li>
      <rdf:li>control</rdf:li>
      <rdf:li>person</rdf:li>
      <rdf:li>aux</rdf:li>
   </rdf:Seq>
</owl:oneOf>
</rdfs:Class>

<rdfs:Class                 rdf:about="http://www.cs.vassar.edu/~ide/rdf/isle-enumerated-
classes#SynFeatureValue">
<owl:oneOf>
   <rdf:Seq>
      <rdf:li>have</rdf:li>
      <rdf:li>be</rdf:li>
      <rdf:li>subject_control</rdf:li>
      <rdf:li>object_control</rdf:li>
      <rdf:li>masculine</rdf:li>
      <rdf:li>feminine</rdf:li>
   </rdf:Seq>
</owl:oneOf>
</rdfs:Class>

</rdf:RDF>
```

Sample LDCR entry for two **Phrase** objects

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:mlc="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#">

<Phrase rdf:ID="NP" rdfs:label="NP"/>

<Phrase rdf:ID="Vauxhave">
   <hasSynFeature>
     <SynFeature>
        <hasSynFeatureName rdf:value="aux"/>
        <hasSynFeatureValue rdf:value="have"/>
      </SynFeature>
   </hasSynFeature>
</Phrase>

</rdf:RDF>
```

Sample LDCR entry for a **Construction** object

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
          xmlns="http://www.cs.vassar.edu/~ide/rdf/isle-schema-v.6#">

<Construction rdf:ID="TransIntrans">
    <slot>
       <SlotRealization rdf:ID="NPsubj">
          <hasFunction rdf:value="Subj"/>
          <filledBy                 rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-
datcats/Phrases#NP"/>
       </SlotRealization>
    </slot>
    <slot>
       <SlotRealization rdf:ID="NPobj">
          <hasFunction rdf:value="Obj"/>
          <filledBy                 rdf:resource="http://www.cs.vassar.edu/~ide/rdf/isle-
datcats/Phrases#NP"/>
       </SlotRealization>
    </slot>
</Construction>

</rdf:RDF>
```

Appendix D: The SIMPLE Ontology

General Ontology for Nouns and Verbs

1. TELIC **[Top]**

2. AGENTIVE **[Top]**

2.1. CAUSE **[Agentive]**

3. CONSTITUTIVE **[Top]**

3.1 PART **[Constitutive]**

3.1.1. BODY_PART **[Part]**

3.2. GROUP **[Constitutive]**

3.2.1. HUMAN_GROUP **[Group]**

3.3. AMOUNT **[Constitutive]**

4. ENTITY **[Top]**

4.1 CONCRETE_ENTITY **[Entity]**

4.1.1 LOCATION **[Concrete_entity]**

4.1.1.1. *3_D_location* **[Location]**

4.1.1.2. *Geopolitical_location* **[Location]**

4.1.1.3. *Area* **[Location]**

4.1.1.4. *Opening* **[Location | Agentive]**

4.1.1.5. *Building* **[Location | Artifact$_{\text{Agentive}}$ | Telic]**

4.1.1.6. *Artifactual_area* **[Location | Artifact$_{\text{Agentive}}$ | Telic]**

🔔 *recommended*

4.1.2. MATERIAL **[Concrete_entity | Telic]**

4.1.3. ARTIFACT **[Concrete_entity | Agentive | Telic]**

4.1.3.1. *Artifactual_material* **[Concrete_entity | Artifact$_{\text{Agentive}}$ | Material$_{\text{Telic}}$]**

4.1.3.2. *Furniture* **[Concrete_entity | Artifact$_{\text{Agentive}}$ | Telic]**

4.1.3.3. *Clothing* **[Concrete_entity | Artifact$_{\text{Agentive}}$ | Telic]**

4.1.3.4. *Container* **[Concrete_entity | Artifact$_{\text{Agentive}}$| Telic]**

4.1.3.5. *Artwork* **[Concrete_entity | Artifact$_{\text{Agentive}}$]**

4.1.3.6. *Instrument* **[Concrete_entity | Artifact$_{\text{Agentive}}$ | Telic]**

4.1.3.7. Money **[Concrete_entity | Artifact$_{\text{Agentive}}$ | Telic]**

4.1.3.8. Vehicle **[Concrete_entity | Artifact**$_{\textbf{Agentive}}$ **| Telic]**

4.1.3.9. Semiotic_artifact **[Concrete_entity | Artifact**$_{\textbf{Agentive}}$**| Telic]**

4.1.4. FOOD **[Concrete_Entity| Telic]**

4.1.4.1. *Artifact_Food* **[Concrete_entity | Artifact**$_{\textbf{Agentive}}$ **| Food**$_{\textbf{Telic}}$**]**

🔔 *recommended*

4.1.4.2. *Flavouring* **[Concrete_entity | Food**$_{\textbf{Telic}}$**]**

🔔 *recommended*

4.1.5. PHYSICAL_OBJECT **[Concrete_entity]0**

4.1.6. ORGANIC_OBJECT **[Concrete_entity]**

4.1.7. LIVING_ENTITY **[Concrete_entity]**

4.1.7.1. *Animal* **[Living_entity]**

4.1.7.1.1. Earth_animal **[Animal]** 🔔 *recommended*

4.1.7.1.2. Air_animal **[Animal]** 🔔 *recommended*

4.1.7.1.3. Water_animal **[Animal]** 🔔 *recommended*

4.1.7.2. *Human* **[Living_entity]**

4.1.7.2.1. People **[Human]**

4.1.7.2.2. Role **[Human]**

4.1.7.2.2.1 Ideo **[Role]**

4.1.7.2.2.2 Kinship **[Role]**

4.1.7.2.2.3 Social_status **[Role]**

4.1.7.2.3. Agent_of_temporary_activity **[Human | Agentive]**

4.1.7.2.4. Agent_of_persistent_activity **[Human | Telic]**

4.1.7.2.5. Profession **[Human | Telic]**

4.1.7.3. *Vegetal_entity* **[Living_entity]**

4.1.7.3.1. Plant **[Vegetal_entity]**

4.1.7.3.2. Flower **[Vegetal_entity]**

4.1.7.3.3. Fruit **[Vegetal_entity]**

4.1.7.4. Micro-organism **[Living_entity]**

4.1.8. SUBSTANCE **[Concrete_entity]**

4.1.8.1. *Natural_substance* **[Substance]**

4.1.8.2. *Substance_food* **[Substance | Food**$_{\textbf{Telic}}$**]** 🔔 *recommended*

4.1.8.3. *Drink* **[Substance | Telic]** 🔔 *recommended*

4.1.8.3.1 Artifactual_drink **[Substance | Artifact$_{\mathbf{Agentive}}$ | Drink$_{\mathbf{Telic}}$]** 🔔 *recommended*

4.2. **PROPERTY** **[Entity]**

4.2.1. QUALITY **[Property]**

4.2.2. PSYCH_PROPERTY **[Property]**

4.2.3. PHYSICAL_PROPERTY **[Property]**

4.2.3.1. *Physical_power* **[Physical_property]** 🔔 *recommended*

4.2.3.2. *Color* **[Physical_property]** 🔔 *recommended*

4.2.3.3. *Shape* **[Physical_property]** 🔔 *recommended*

4.2.4. SOCIAL_PROPERTY **[Property | Agentive]** 🔔 *recommended*

4.3. **ABSTRACT_ENTITY** **[Entity]**

4.3.1. DOMAIN **[Abstract_entity]**

4.3.2. TIME **[Abstract_entity]**

4.3.3. MORAL_STANDARDS **[Abstract_entity]** 🔔 *recommended*

4.3.4. COGNITIVE_FACT **[Abstract_entity | Agentive]**

4.3.5. MOVEMENT_OF_THOUGHT **[Abstract_entity | Agentive]**

4.3.6. INSTITUTION **[Abstract_entity | Agentive | Telic]**

4.3.7. CONVENTION **[Abstract_entity | Agentive]** 🔔 *recommended*

4.4. **REPRESENTATION** **[Entity | Agentive | Telic]**

4.4.1. LANGUAGE **[Representation]**

4.4.2. SIGN **[Representation]**

4.4.3. INFORMATION **[Representation]**

4.4.4. NUMBER **[Representation]** 🔔 *recommended*

4.4.5. UNIT_OF_MEASUREMENT **[Representation]**

4.5. **EVENT** **[Entity]**

4.5.1. PHENOMENON **[Event]**

4.5.1.1. *Weather_verbs* **[Phenomenon]** 🔔 *recommended*

4.5.1.2. *Disease* **[Phenomenon | Agentive]** 🔔 *recommended*

4.5.1.3. *Stimuli* **[Phenomenon | Agentive]** 🔔 *recommended*

4.5.2. ASPECTUAL **[Event]**

4.5.2.1. *Cause_aspectual* **[Aspectual | Cause$_{\mathbf{Agentive}}$]**

4.5.3. STATE (event type=***state***) **[Event]**

4.5.3.1. *Exist* **[State]**

4.5.3.2. *Relational_state* **[State]**

4.5.3.2.1. Identificational_state **[Relational_state]** 🔔 *recommended*

4.5.3.2.2. Constitutive_state **[Relational_state]** 🔔 *recommended*

4.5.3.2.3. Stative_location **[Relational_state]** 🔔 *recommended*

4.5.3.2.4. Stative_possession **[Relational_state]** 🔔 *recommended*

4.5.4. Act **[Event]** (event type=***process***)

4.5.4.1. *Non_relational_act* **[Act]**

4.5.4.2. *Relational_act* **[Act]**

4.5.4.2.1. Cooperative_activity **[Relational_act | Agentive]** 🔔 *recommended*

4.5.4.2.2. Purpose_act **[Relational_act | Telic]** 🔔 *recommended*

4.5.4.3. *Move* **[Act]**

4.5.4.3.1 Caused_motion **[Move | Cause$_{Agentive}$]**

4.5.4.4. *Cause_act* **[Act | Cause$_{Agentive}$]**

4.5.4.5. *Speech_act* **[Act]**

4.5.4.5.1. Cooperative_speech_act **[Speech_Act]** 🔔 *recommended*

4.5.4.5.2. Reporting_events **[Speech_Act | Telic]** 🔔 *recommended*

4.5.4.5.3. Commissives **[Speech_Act | Telic]** 🔔 *recommended*

4.5.4.5.4. Directives **[Speech_Act | Telic]** 🔔 *recommended*

4.5.4.5.5. Expressives **[Speech_Act | Telic]** 🔔 *recommended*

4.5.4.5.6. Declaratives **[Speech_Act | Telic]** 🔔 *recommended*

4.5.5. Psychological_event **[Event]**

4.5.5.1. *Cognitive_event* **[Psychological_event]**

4.5.5.1.1. Judgment **[Cognitive_event | Telic]** 🔔 *recommended*

4.5.5.2. *Experience_event* **[Psychological_event | Agentive]**

4.5.5.2.1. Caused_Experience_event **[Experience_event | Cause$_{Agentive}$]**

4.5.5.3. *Perception* **[Psychological_event]**

4.5.5.4. *Modal_event* **[Psychological_event | Telic]**

4.5.6. CHANGE **[Event]** (event type=***transition***)

4.5.6.1. *Relational_change* **[Change | Agentive]**

4.5.6.1.1. Constitutive_change **[Relational_change | Agentive]**
🔔 *recommended*

4.5.6.1.2. Change_of_state **[Relational_change | Agentive]**
🔔 *recommended*

4.5.6.1.3. Change_of_value **[Relational_change | Agentive]**
🔔 *recommended*

4.5.6.2. *Change_possession* **[Change | Agentive]**

4.5.6.2.1. Transaction **[Change_possession]**

4.5.6.3. *Change_of_location* **[Change | Agentive]**

4.5.6.4. *Natural_transition* **[Change| Agentive]**

4.5.6.5. *Acquire_knoweldge* **[Change| Agentive]**

4.5.7. CAUSE_CHANGE **[Event | Cause$_{\text{Agentive}}$]**

4.5.7.1. *Cause_relational_change* **[Cause_change]**

4.5.7.1.1. Cause_constitutive_change
[Cause_Relational_change] 🔔 *recommended*

4.5.7.1.2. Cause_change_of_state
[Cause_Relational_change] 🔔 *recommended*

4.5.7.1.3. Cause_change_of_value
[Cause_Relational_change] 🔔 *recommended*

4.5.7.2. *Cause_ change_location* **[Cause_Change]**

4.5.7.3. *Cause_ natural_transition* **[Cause_Change]**

4.5.7.4. *Creation* **[Cause_Change]**

4.5.7.4.1. Physical_creation **[Creation]** 🔔 *recommended*

4.5.7.4.2. Mental_creation **[Creation]** 🔔 *recommended*

4.5.7.4.3. Symbolic_creation **[Creation]** 🔔 *recommended*

4.5.7.4.4. Copy_creation **[Creation]** 🔔 *recommended*

4.5.7.5. *Give_knoweldge* **[Cause_Change | Telic]**

General Ontology for Adjectives

1. **INTENSIONAL** **[Top]**

1.2. **Modal** **[Intensional]**

1.3. **Temporal** **[Intensional]**

1.4. **Emotive** **[Intensional]**

1.5. **Manner** **[Intensional]**

1.6. **Object-related** **[Intensional]**

1.7. **Emphasizer** **[Intensional]**

2. **EXTENSIONAL** **[Top]**

2.1. **Physical_property** **[Extensional]**

2.2. **Psychological_property** **[Extensional]**

2.3. **Social_property** **[Extensional]**

2.4. **Temporal_property** **[Extensional]**

2.5. **Intensifying_property** **[Extensional]**

2.6. **Relational_property** **[Extensional]**

Appendix E: EuroWordNet Top Ontology

Top^{0}

1stOrderEntity^{1}

- Origin$^{\mathbf{0}}$
  - Natural^{21}
    - Living^{30}
      - Plant^{18}
      - Human^{106}
      - Creature^{2}
      - Anima^{l23}
  - Artifact^{144}
- Form$^{\mathbf{0}}$
  - Substance^{32}
    - Solid^{63}
    - Liquid^{13}
    - Gas^{1}
  - Object1^{62}
- **Composition**^{0}
  - Part^{86}
  - Group^{63}
- **Function**^{55}
  - Vehicle^{8}
  - Representation^{12}
    - MoneyRepresentation^{10}
  - LanguageRepresentation^{34}
    - ImageRepresentation^{9}
  - Software^{4}
  - Place^{45}
  - Occupation^{23}
  - Instrument^{18}
  - Garment^{3}
  - Furniture^{6}
  - Covering^{8}
  - Container^{12}
  - Comestible^{32}
  - Building^{13}

2ndOrderEntity^{0}

- SituationType$^{\mathbf{6}}$
  - Dynamic^{134}
    - BoundedEvent^{183}
    - UnboundedEvent^{48}
  - Static^{28}
    - Property^{61}
    - Relation^{38}
- SituationComponent$^{\mathbf{0}}$
  - Cause^{67}
    - Agentive^{170}
    - Phenomenal^{17}
    - Stimulating^{25}
  - Communication^{50}
  - Condition^{62}
  - Existence^{27}
  - Experience^{43}
  - Location^{76}
  - Manner^{21}
  - Mental^{90}
  - Modal^{10}
  - Physical^{140}
  - Possession^{23}
  - Purpose^{137}
  - Quantity^{39}
  - Social^{102}
  - Time^{24}
  - Usage^{8}

3rdOrderEntity$^{\mathbf{33}}$

Appendix F: Representing noun compounds and support verbs in MILE (PISA & XMELLT)

This section contains a general overview of the state of the art regarding MWEs, with particular focus on Complex Nominals and Support Verb Constructions, in a multilingual perspective. Two European languages, namely English and Italian, have been chosen as main sources from which to draw examples: because they adopt quite different syntactic strategies, they prove to be quite interesting for a cross-lingual study .
It follows a case study of few subtypes of MWEs, for a possible lexical semantic representation in MILE.
The data reported in this section have all been extracted from two representative corpora (where not differently specified): the PAROLE corpus for the Italian language and the British National Corpus World-Edition for English (Burnard, 2001).
The main focus of such case studies has been a parallel analysis in the two languages, in order to discover what are the common and what the different information one needs in a multipurpose, multi-language computational lexicon.
Some hints at the German and/or French equivalents will sometimes also be given, but not supported with corpus data.

F.1 Multi-Word Expressions (MWEs): a challenge to NLP

MWEs are a class of syntactically (and semantically) complex lexical units, i.e. expressions composed of more than one lexical item that should nevertheless be recognized as a unit at the lexical, semantic or conceptual level.
MultiWord Units (MWUs) include a variety of expressions showing different degrees of compositionality and different degrees of discontinuity: the range goes from almost fully compositional syntactic patterns to fully idiomatic (i.e. fixed) expressions.
Below it is given a possible classification of MWUs, according to their degree of semantic transparency, syntactic discontinuity and to some other relevant features (Atkins Personal Communications, 2002):

1. Fixed/ semi-fixed phrases:
   - a. irreversible pairs/triples: *ham and eggs, fish and chips…*
   - b. transparent similies: *white as a sheet, pale as death…*
   - c. catch phrases: *if you can' t beat ' em, join ' em, horses for courses…*
   - d. proverbs: *too many cook' s spoil the broth, birds of a feather…*
   - e. quotation: *to be or not to be, eye for an eye…*
   - f. greetings: *good morning, how do you do?*
   - g. phatic phrases: *have a nice day, take care of yourself*

2. Non-compositional compounds:

   (The following are some criteria useful to identify them: they must be fixed multiword expressions, must participate in semantic relationships (synonymy, antonym etc.) with single words, often have single-word translation in another language).

   It is possible to sub-classify noun compounds into the following subtypes, according to the semantic relation the expression hold with its head noun, and/or to the sense of its component –or one of them:

a. figurative compounds: (are no hyponym of their heads, or do not necessarily have a similar denotation of the heads): Ex. *lame duck, civil servant…*

b. semi-figurative compounds: *high school, sky blue…*(most of these are ADJ + N compounds, they are not fully compositional in that the compound is a hyponym of its head but cannot be paraphrased as a HEAD which is ADJ. Ex. *A high school* is not "a school which is high").

c. functional compounds: (the compound is a HEAD that has to do with the thing denoted by the modifier, but also more than that: it is a specific type of thing or person, denoted by the HEAD.): *house agent, police dog, can opener, level crossing…*

3. Idioms:
   a. lexically inflexible : *by and large, hand over fist…*
   b. lexically flexible :
      -lexical alternation: *to throw in the sponge (*or *towel)*
      - variability: *chicken and egg (which came first, the chicken or the egg?, it' s a chicken-and-egg situation…)*
      - lexical gapping: *it was a …' s dream*(slot: activity-linked noun)
      - semantic alternation: *to have a heart of gold/ to have a heart of stone*
   c. syntactically inflexible (limited grammatical transformations): *it was a football manager' s dream / * the dream of a football manager.*
   d. morpho-syntactic flexibility: verb tense and agreement of possessives *to get too big for one' s booths.*
4. Support-verb constructions: *take a walk, have a bath, make a decision…*

5. Phrasal Verbs:
   - V+adv: *get up…*
   - V+P: *break into…*
   - V+adv+P: *come up with…*

What all these cases seem to share is the quite general fact that *a sequence of words acts as a single unit at some level of linguistic analysis*. These types of expressions also share some of the following general features, usually brought as evidence for a given sequence of words to form a multiword element:

1. reduced syntactic and semantic transparency;
2. reduced or lack of compositionality;
3.possible violation of some otherwise general syntactic pattern or rules;
4. high degree of lexicalization (depending on pragmatic factors);
5. high degree of conventionality.

MWEs differ for the degree in which those features occur, making the range spanning from full-fledged compositional and productive constructions to fixed idioms (Calzolari&Lenci "ISLE Preliminary Notes"). However, because different parameters can be taken at a time in order to define the different classes, it is difficult to treat MWEs as a whole. Each type should be appropriately defined before investigating it, but in most cases standard definitions have not been adopted.

Among the types of MWEs, Complex Nominals and Support Verb Constructions seem to be the most interesting as long as they are the most hard to describe and difficult to treat computationally. They represent the *hard cases* in the realm of MWEs, since they call for a different status with respect to purely compositional constituents, and show an internal cohesion and an extremely high degree of variability in lexicalization and language-dependent variation.

In computational approaches MWEs are those constructions that cannot be properly understood if they are not recognized as complex lexical items -and thus to be entered in the lexicon as separate units-, i.e. those expressions that cannot fully be treated by the standard rules of the grammar.

From a multilingual perspective, the treatment of MWEs is even harder: any attempts to establish links between language equivalents of MWEs seems to require simultaneous access to various levels of description. In fact, literal translations are often not acceptable, and quite often there is no one-to-one matching between MWEs of different languages.

Even when a one-to-one mapping does exist, it is very frequent that their syntactic structures differ considerably: Ex. Eng. *typewriter* = It. *Macchina da scrivere*.

Or different lexical choices must be made: Eng. *Take a shower* = it. *Fare una doccia*

F.2 Complex nominals

F.2.1 CN in a theoretical perspective

Complex Nominals are NPs which happens to be not fully regular at one level of linguistic analysis, esp. at the syntactic and/or semantic levels. The fact that the CN class is not homogenous –i.e. the various instances of CNs show different degrees of idiosyncrasies and different syntactic structure esp. when considered from a multilingual perspective- makes it difficult to establish useful criteria to treat the class as a whole.

Typical English representatives are the well known and investigated expressions traditionally called *noun compounds*, that is nouns pre-modified by a noun, adjective, possessive phrase or gerund.

Apart from the problems posed to the lexicographer by lexicalized compounds, this kind of constructions has been studied theoretically mostly to discover the regular patterns responsible for the formation of the so called *novel compounds*, i.e. compounds formed anew in a particular situation, in analogy with already existing compounds or semantically motivated by the context. The problem of novel compounds does not belong to the lexicon, but an appropriate analysis/treatment of lexicalized compounds could give other components the patterns by which those formations can be interpreted as well. One of the greatest difficulties in interpreting even novel compounds consist in the retrieval of the implicit semantic relation linking the formatives in a compound, in order to have a full interpretation of the construction. The appropriate place for such information to be stored appears to be the lexicon, given also the fact that novel compound formation seems to be mostly an analogical process based on already existing and well-established compounds (see Lyons, 1977; Downing, 1977; Warren, 1978).

As previously mentioned, language variation is high. In other languages, like Italian or French, CNs are NPs correspond to a head N plus a post-modifier element: an adjective, an infinitive clause or a Prepositional Phrase (PP_*di*, PP_*da* or PP_*a*); in such constructions no lexical item can normally intervene between the head N and the PP, and the post-modifier noun usually occur without a determiner:

[1] Ex. bicchiere da vino ('wine glass')*, carta di credito* ('credit card')*, barca a vela* ('sailing boat')

One of the first problems from a computational perspective is the recognition of CNs, distinguishing them from their completely regular syntactic counterparts.

F.2.2 ENGLISH COMPOUND NOUNS

In linguistic theory *noun compounds* have been studied from the phonological, morphological, syntactic and/or semantic point of view in order to discover their defining characteristics. In fact, all came out with a set of diagnostics and tendencies that can help recognize them, but cannot have an heuristic value.

Among the most significant studies it is worth citing Downing (1977), Levi (1978), Warren (1978), Leonard (1986), Chomsky and Halle (1968).

Moreover, it must be highlighted that most diagnostics are language dependent: i.e. each language has its own peculiar strategies for CN formation. However, some diagnostics have been identified which work both for English and Italian.

F.2.2.1. PHONOLOGICAL CHARACTERISTICS

The following characteristics apply to English only, phonological properties of lexical items being strongly language dependent.

Traditionally, *noun compounds* are those NPs (usu. N+N) that do not follow the normal unmarked stress pattern of English phrases: in regular NPs the primary stress is carried by the head noun, that is the rightmost element. *Noun compounds,* on the other hand, usually have word-stress: primary stress falls on the first element, the modifier:

[2] Black bird vs. Blackbird (Chomsky & Halle, 1968, 93).

This criterion has long been taken as the only viable way to distinguish *noun compounds* from regular syntactic constructions. However, there is much debate on the topic; empirical data have demonstrated a high degree of variability in stress pattern in real use, even of acknowledged compounds (cfr. Pennanen 1980).

F.2.2.2 ORTHOGRAPHIC CHARACTERISTICS

Orthographic characteristics of CNs are highly language dependent, orthography being fundamentally a matter of convention.

English *noun compounds* show a great variability in orthographic realizations: many compounds are realized as one word, i.e. as a continuous string of characters: *typewriter, ashtray, blackboard…*

These are the most typical realizations of compounds, for orthography generally reflects word-stress pattern. Such compounds are perceived as single bases by native speakers. According to Warren (1978), this fact might be taken as a mark of their being totally lexicalized and, therefore, to be entered into the lexicon as single lexical items. She moreover observes that novel compounds are hardly ever realized as one word.

Nevertheless, in English there are two other kinds of very frequent realizations: compounds can be realized as two words linked by an hyphen or separated by a space.

According to Warren, the use of hyphens is a mark of the perceived syntactic-semantic unity of the sequence: Ex. *worker-bee.*

Compounds realized this way are usually lexicalized or at least *institutional,* i.e. very frequent collocations and category-denoting expressions[26].

Finally, most *compound nouns* are realized as two juxtaposed words, i.e. like NPs (Ex. *toy knife, apple cake*). They are normally novel compounds, but it happens to be not infrequent to have corpus attestations of institutionalized, lexicalized and even idiomatic compounds realized this way (ex. *Melting pot, black box…*).

[26] Occasionally, however, it is possible also for a novel compound to be realized with a hyphen separating its formatives.

F.2.2.3 MORPHOLOGICAL COMPOSITION.

At the morphological level we can identify nine types of English noun compounds:

1- N+N; *air rifle*
2- Adj.+N; *musical clock*
3- N' s+N; *cook' s knife*
4- V(stem)+N; *pickpocket*
5- Ver+N; *worker-bee*
6- N+Ver; *pencil sharpener*
7- Ving+N; *hunting rifle*
8- N+Ving; *net surfing*
9- Ved+N; *lidded box*

These nine types are all the possible types of *noun compounds* that can be found in English, although some of them is scarcely productive. Many linguists –cf. inter alia Quirk et al. (1985) and Warren (1978)- have tried to formulate possible tendencies or preferences of compound formation in English, given that it seems still not to be possible to fix rules because of the great number of exceptions in the morphological composition of compounds.

In particular, they observe that generally the noun modifier behaves differently w.r.t its normal use: it cannot be preceded by a determiner (any such item will determine either the head noun or the whole compound); it cannot be inflected (but here there are many counterexamples) nor modified by an adjective unless it is in its turn compounded with it.

The normal occurrence of the modifier noun in the base form poses problems of semantic ambiguity: in the process of interpretation it is up to the receiver (listener or reader) to determine whether the noun refers to a singular or plural entity on the basis of the context or of other kinds of knowledge.

There are, however, many examples of compounds with a plural noun modifier: some of them are taken as if they were regular since the modifier happens to be an *exclusive plural*, that is a nouns which has been lexicalized at the plural form only. In other cases, we must talk about of counter-examples: these are nevertheless very informative from the semantic perspective because they prevent a sort of unresolvable ambiguity; some times alternative compounds exist and can even be lexicalized with different meanings: ex. *Career girl* vs. *careers girl*.

As far as compounds with a modifier inflected for the genitive are concerned, Warren notes their potential ambiguity with regular NP expressing possession. Following Quirk, it seems clear that the ambiguity can be resolved only on semantic grounds with a set of selection preferences for certain semantic kinds of modifiers: for N' s +N constructions to be compounds, according to Quirk the modifier must respect the following conditions: first of all, it must be a restrictive genitive and then it must denote a social status: *a fisherman' s cottage* vs. *a friend' s cottage*.

F.2.2.4 SYNTACTIC CHARACTERISTICS

Compounds realized as one word or with their elements united by a hyphen are usually treated like other simple lexical items–although this is a debatable matter- and therefore do not receive syntactic analysis.

At the syntactic level, the most interesting types of *noun compounds* in English are those realized as two words, which must be assigned both a lexical description and a syntactic structure.

Being NPs, compounds are traditionally classified into two major classes: endocentric and exocentric compounds (some linguists adds two more classes: copulative and appositional, but the constructions belonging to these could well be included in the two mentioned above).

Endocentric compounds have the same syntactic distribution of their heads:

[3] bread knife → knife, coltello da pane → coltello.

Exocentric compounds, instead, do not share the same distributions with their heads:

[4] ladybird ⊄ bird, red cap ⊄ cap…

Exocentric compounds are not investigated because they often involve metaphoric extension and, thus, are to be included within investigations on metaphor, a process which involves mechanisms which are different – though perhaps complementary- from those of compounding.

F.2.2.5 SEMANTIC COMPOSITION

The fundamental peculiarity of *noun compounds* at the semantic level is that they constitute single conceptual/denotational units, they are felt as building a new conceptual unit of their own (Quirk 95).

Their lexical meaning, however, can be, at varying degree, reconstructed (except for fully idiomatic expressions) by the meanings of their components plus the underlying semantic relation linking them.

In endocentric compounds, those almost exclusively investigated, the modifier adds a semantic content which serves to restrict the set of referents of the head noun specifying how the designatum is a subtype of them, so that the whole unit turns out to be a hyponym of its head noun. (Classificatory Compounds)

[5] Ex. Cruise ship (is a kind of ship).

Or to indicate which is the instance in focus (Deictic Compounds)

[6] Ex. The orange-juice seat ('the seat in front of which there is a glass of orange juice'; ex. Quoted from Downing [1977].).

At the same time, however, the modifier looses part of its own referential potential:

[7] A wine glass (un bicchiere da vino).
is not necessarily full of wine. Therefore, *wine* here has a reduced referential power.

The problem of the semantics of noun compounds and CNs in general is not an easy topic, since it is still unclear how much/what should be treated within a theory of the lexicon and what concerns conceptual organization in a KB, inference rules and pragmatics.

What appears to be necessary in a lexicon is to make available the semantic relationships underlying compounds, a kind of information which proves indispensable either in MT, IE etc..

F.2.2.6 SEMANTIC RELATIONSHIPS IN COMPOUNDS

Downing (1977) believes -in contrast with the generative tradition- that the number of the semantic relations that can underlie compounds is not finite, or in any case not determinable *a priori.* She observes that the context and extra-linguistic knowledge of the hearer are factors which strongly lead the speaker/hearer ability of creating/interpreting noun compounds. Restrictions, if there exist, must be of a pragmatic rather than of a semantic nature. She claims the impossibility of formalizing all factors determining the range of situations in which a certain compound can be used in terms of a finite list of all possible relations holding between formatives. She believes, in fact, that every semantic relation can virtually function well in an adequate context.

Nevertheless, she admits that attested relations can be reduced to a set of elementary predicates- like those proposed by Levi (1978)- but at the cost of a considerable loss of semantic material. It is possible then to identify a limited number of underlying structures, formed with one of those primitive predicates, from which general classes of compounds might be derived. Such structures, however, will never be exactly equivalent to the precise relation linking the terms of the compound, since general structures cannot describe all semantic nuances existing in the language.

Such a reduction, however, can be practically useful: it has been observed, in fact, that certain relations have a higher classificatory power w.r.t. others and are more frequent in discourse. The identification of these -more relevant and more frequent- relations would, then, allow for the construction of models of the language accounting at least for a high number of compounds -if not for all possible ones.

The most important result, of interest here, emerging from Downing' s work is that the types of relations underlying compounds strongly depend on 3 factors:

1- the semantic class of the head noun: certain classes of nouns seem to select preferably for certain relations, while other relations seem less relevant.

2 - The predictability of the relation: the more predictable -i.e. general or permanent- a relation is, the more easily it can be interpreted and the higher is its portability in other context. Totally predictable relations, however, are not suitable for compounding, except for in special cases – lexicalized compounds (ex. *lime tree*)- or contexts, because the information brought by the modifier would be redundant.

3 - the permanence of the relation: most frequently used compounds are based on habitual or generic relations, or on relations indicating some inherent characteristics of the head noun.

Levi' s theory (1978) aims at modeling the productive mechanisms of nominal composition; her final goal is, thus, to formulate general (grammatical) rules which account for the generation of all possible interpretations of regular productive CNs, leaving it to other linguistic component to fix the conditions under which lexical items may be combined.

According to this model, there are only two possible mechanisms responsible for the generation of CNs: *Predicate Deletion* and *Predicate Nominalization*; PD occurs when the predicate which is present in the deep structure "gets lost" in the surface structure. This mechanism gives rise to CNs of the N+N and Adj.+N types.

Levi claims there exists only 9 semantically primitive predicates, called *Recoverably Deletable Predicates* (RDPs), representing all and only those semantic relations that can be left unexpressed in English CNs.

These predicates are: *CAUSE, HAVE, BE, MAKE, USE, IN, FOR, FROM, ABOUT,* and express the semantic relationships that are traditionally identified as: causative, possessive/ dative, productive/ constitutive, instrumental, appositional, locative, purposive/ benefactive, source/ ablative, topic.

Her theory, however, does not explain how the 9 RDPs can be recovered in the process of interpretation/ decoding. It is not at all clear, that is, how the listener/reader can perform the inverse operations leading from the CN to its possible deep structures, and thus to its possible meanings/interpretations.

Warren (1978): From her corpus study, she obtains a hierarchy with 6 top nodes, i.e. 6 major classes:

CONSTITUTE: *claybird, student group;*
POSSESSION: *board member, apple pie;*
LOCATION: *moon rocket, weekend guest;*
PURPOSE: *table-cloth;*
ACTIVITY-ACTOR: *room clerk;*
RESEMBLANCE: *bullethead.*

These classes are in their turn subdivided into other subtypes. Altogether Warren identifies 54 relations that can be left unexpressed in the construction of noun sequences (or non-verbal compounds). These relations, except for idiomatic sequences and copulative compounds, can be paraphrased by PPs in which they appear in an explicit way, i.e. they are syntactically realized.

F.2.3 ITALIAN COMPLEX NOMINALS

In Italian CNs formation exploits post-modification; here we have three types of CNs: N+N (*nave scuola, capostazione*), N+Adj (*coltello elettrico*), N+PP (*coltello da pane*). Compounds in Italian are not very frequent and mostly lexicalized, they are often foreign claques.

Nouns post-modified by restrictive adjectives are to be considered CNs in that the adjective functions like a noun. Such adjectives are similar to the *non predicative* adjectives described in English linguistic literature.

Of greater interest in the field of MWEs is the kind N+PP, the one that will be considered more extensively here. This pattern of CNs formation is highly productive in Italian, and in romance languages in general, and such CNs very often translate compound nouns of other -mainly Germanic- languages.

F.2.3.1 MORPHOSYNTACTIC CHARACTERISTICS

At the morphological level the only significant peculiarity of N+PP complex nominals is that the modifier usually occur only either in the singular or in the plural form, depending on the type or even on the particular CN. This appears to be purely a lexical choice.

From the strictly syntactic point of view, it can be noticed that generally only three prepositions occur in such CNs: that is *a, di* and *da*; and no element can normally intervene between the elements of the construction.

The noun in the modifier PP, moreover, tends to have no determiner, but this does not constitute a rule.

F.2.3.2 SEMANTIC PROPERTIES

CNs of the form N+PP seem to share most of the semantic characteristic of compound nouns, presented above.

The presence of a preposition, however, is taken to be an explicit mark of the semantic relation underlying the compound (Johnston&Busa 1999:169), though not very precise.

F.2.3.3 SUMMARY

It seems clear that a more detailed classification of CNs at different levels is not only useful but necessary. Such classifications could be based on the various peculiarities detected at the different levels of description, but would be language dependent.

A general, not necessarily language dependent, syntactic distinction to be drawn is between N+N and N+PP types (ex. *chest of drawers* vs. *cruise ship*; *nave scuola* vs. *vaso da notte*).

English N+N can be classified on the basis of the morphological type of modifier, and/or of the head.

N+PP, instead can be distinguished according to the kind of the occurring preposition.

From the semantic perspective, moreover, both types can be classified on the basis of the productive relation linking the two formative, but to do that a detailed, predetermined hierarchy of relations must be available.

Semantically, it is also possible to classify CNs taking into account the status of the senses of the formatives: whether they are used in one of their common senses, in a figurative sense or only in that specific/some limited contexts. (see Lyons 77:544).

F.2.4 POSSIBLE DIAGNOSTICS FOR THE IDENTIFICATION OF CNs

Some diagnostics that may help identify what are to be considered CNs are indeed needed. Nevertheless, it must be stressed that such tests are never completely discriminating nor decisive.

A general diagnostic could be their capability of entering in semantic relationships -such as synonymy, antonymy or hyponymy- with single lexical items.

Moreover, what seems to be really stringent is to be able to provide a satisfactory explicit characterization of the internal structure of the constructions.

1- reduced or lack of *semantic referentiality* of the modifier;

2- CNs must define a new subtype of the entity denoted by the head of the construction;

3- morphosyntactic clues: impossibility of internal modification (except in cases of nested CNs);

- in English compounds: occurrence of the modifier in the singular form ;
- in Italian CNs: lack of determiner of the modifier noun in the PP.

F.2.5 CNs FROM A COMPUTATIONAL POINT OF VIEW

Complex nominals play an important role in language - esp. in the expression of (new) nominal concepts - and are frequent in a great variety of texts: medical, journalistic, technical etc. Therefore, the capability of handling them by NLP systems is essential for new generation applications, esp. for MT and IR/IE.

At present, few systems have tried to deal with CNs, and the results are not very satisfactory. In such systems compounds have generally received a compositional treatment, and therefore non-compositional, lexicalized or "problematic" cases have been disregarded, attributed to the lexicon without, however, any further detail. The problem of representation and description is, instead, not an easy task at all.

Johnston & Busa (1999) propose a compositional analysis of English and Italian complex nominals, based on the types of relational information given by the Generative Lexicon model (esp. the Qualia Structure), which would limit the need for listing compound in the lexicon and bring the interpretation/generation of complex nominals under the rubric of other compositional mechanisms in language, as co-composition, and type-coercion. This account, aiming at semi-automatic lexical acquisition of CNs, starting from the representation of single lexical items in an established and already implemented lexion, might even prove useful for the representation also of non-compositional compounds, for which at least some information that cannot be inferred compositionally must be made explicit.

Here, the modifier in a CN would specify either the semantic type of one of the participants in one of the Qualia Roles activated by the nominal head or one of the relations expressed in its Qualia Structure. By making use of phrase structure schemata it is possible to link the lexical semantic representation of items to their syntactic expressions and then to compose them in the compound form.

Concerning Italian CNs, the authors treat the modifying phrase not as an actual PP, but consider the preposition as a bound morpheme indicating which Qualia Role is involved in the composition with the modifier (for further details see. Busa & Johnston 1996 and Johnston & Busa 1999).

This strategy may be improved, or somehow reused, by using SIMPLE' s Extended Qualia Structure. The way the appropriate specific semantic relations are to be selected is, however, not clear.

F.2.5.1 COMPLEX NOMINALS IN ROSETTA AND GLOBALINK

F.2.5.1.1 The Rosetta MT system

The Rosetta MT system is a research prototype translating between Dutch, English and Spanish. It was developed in the Rosetta project carried out at the Philips Research Laboratories. For an extensive description of the approach and the system developed, see Rosetta (1994).

The grammars of the Rosetta system are a special kind of compositional grammars, called M-grammars, consisting of basic expressions and rules. One can derive utterances by recursively applying rules, initially to basic expressions. In order to be able to deal adequately with the complexities of natural language, rules apply to (one or multiple) syntactic trees

called S-trees (short for surface trees) that encode syntactic categories, attribute-value pairs, linear order, and constituent structure. Basic expressions are a special kind of S-tree called a lexical S-tree. Compositional grammars are designed in such a way that the principle of Compositionality of Meaning holds. This principle can be stated as follows:

(1) Compositionality of Meaning: The meaning of an expression is a function of the meaning of its parts and the way they are combined.

This is achieved in compositional grammars because both basic expressions and rules have a meaning. Therefore, the meaning of an utterance can be derived in parallel with the syntactic derivation, and the way the meaning is computed can be represented in a semantic D-tree consisting of unique identifiers for the meanings of the rules and the meanings of the basic expressions.

The principle behind the method used to translate with M-grammars is called the principle of *Compositionality of Translation*:

(2) Compositionality of Translation: Two expressions are each other's translation if they are built up from parts which are each other's translation, by means of rules which are each others translation.

Having the same meaning is a necessary condition for translational equivalence, but factors additional to meaning (e.g. stylistic) might be relevant in determining translational equivalence.

M-grammars of two different languages must be isomorphic, in order for this last proncipl e to apply.

For each derivation tree in G1 there will be an isomorphic derivation tree in G2 that is translationally equivalent, and if the grammars satisfy certain additional restrictions, one can derive translationally equivalent utterances from G1 and G2 in parallel.

A property that follows immediately from the approach and that is crucial for the treatment of MWEs is that an MWE that has a non-compositional meaning must be treated as a basic expression in the grammar: The design does not allow multiple expressions to map to one meaning (and it sharply contrasts here with other approaches, e.g. the Globalink approach).

F.2.5.1.2 The Globalink MT system

The Globalink MT system 1 is a transfer-based MT system. It has a grammar used in analysis and generation, but not for analyzing MWEs. The resolution of MWEs is carried out in transfer, where multiple lexical items may map to a single lexical item in the target language, and vice-versa.

F.2.5.2 FIXED MWES

For reasons of practicality and because the degree of fixedness is one of the most prominent properties od MWEs in general in MT, here Nominal MWEs are subclassified into three types, according to their degree of compositionality and syntactic variability.

The simplest types of MWEs are fixed MWEs. Fixed MWEs in Rosetta consist of a sequence of words where:

- the individual words occur in a fixed order
- the individual words are always contiguous (no other elements can intervene)
- there is no variation in lexical item choice
- there is no inflection or only inflection at one edge[27]

Typical examples are fixed expressions and foreign geographic and other names that consist of multiple words:

[8] a. *ad hoc* 'ad hoc', *stante pede 'stante pede', ter plaatse* 'on the spot'*, by and large*

[9] *b. Hong Kong, Kuala Lumpur, New York, San Francisco*

but certain compounds and phrases in languages with little or no inflection (e.g English) could be dealt with as fixed MWEs as well[28]:

[10] credit card, travel agency, real estate agency

Examples of MWEs that cannot be treated as fixed MWEs are given in (5):

[11] a. (En.) *mother-in-law*

[27] There is no principled reason to exclude inflection in the middle, but it is technically easier to deal with if it is excluded and it avoids the use of prefixes and suffixes as infixes, so that the morphological component can remain simple, at least in languages that do not have infixation but probably also in languages with infixation.

[28] That might not be a very principled approach, but it can be very convenient in the development of actual systems.

b. (It.) *carta telefonica* ‘telephone card’
c. (Dutch) *de plaat poetsen* (lit. ‘to polish the plate’, ‘to bolt’)

The English example (5a) cannot be dealt with as a fixed MWE because it has internal inflection (cf. *mothers-in-law*). The Italian example (5b) cannot be treated as such because each of its component words inflects (cf. *carte telefoniche*).

The Dutch example (5c) cannot be treated as a fixed MWE because next to a canonical order with contiguous elements (as in (6a)), it also allows other words to intervene between its components (as in (6b)), it allows permutations of its component words (as in (6c)), and combinations of permutations and intervention by other words not part of the MWE (as in (6d):

[12] a. *Hij heeft gisteren de plaat gepoetst*
lit. ‘He has yesterday the plate polished’
b. *Ik dacht dat hij gisteren de plaat wilde poetsen*
lit. ‘I though that he yesterday the plate wanted polish’
c. *Hij poetste de plaat*
lit. ‘He polished the plate’
d. *Hij poetste gisteren de plaat*
lit. He polished yesterday the plate’

Fixed MWEs are treated in the Rosetta lexicons as normal lexicon entries that require one and allow more than one contiguous spaces in their orthographic representation. In analysis, the incoming sequence of words is mapped onto a single lexical tree by morphology before it enters syntax. This makes it very suitable for dealing with MWEs for which the internal syntax is unclear or irregular relative to the syntax of the system (such as ad hoc, by and large, Dutch op en top ‘fully’, etc.).

F.2.5.2 Semi-Flexible MWEs

The Globalink system allows for (what I will call) semi-flexible MWEs. In this type of MWE,
– the component words have to occur in a fixed order
– the component words have to be contiguous (words that are not part of the MWE cannot intervene)
– more than one part can inflect

Typical examples of such MWEs are given in (7):

[13] a.(En.) *House of Representatives*
b. (Sp.) *patatas fritas*
c. (Fr.) *calculateur analogique*
d. (Fr.) *résistant aux acides*
e. (It.) *carta telefonica*

None of these examples can be dealt with as a fixed MWE because of the internal or multiple inflection (cf. *House**s** of Representatives, patata**s** frita**s**, ca-culateur**s** analogique**s**, résistant**e** aux acides, cart**e** telefoni**che***).

Semi-flexible MWEs are represented in the Globalink lexicon (Lexicon Interchange Format, or LIF) as a sequence of the component words. The head of the phrase is marked as such, and each component word that can be inflected is represented by its canonical form and marked by an asterisk. Typical representations are:

[14] a. (En.) *House* of Representatives*
b. (Sp.) ***patata*** frito**
c. (Fr.) ***calculateur*** analogique**
d. (Fr.) ***résistant*** aux acides*
e. (It.) ***carta*** telefonico**

where the head is marked by putting it in bold face.

In analysis, each word of the incoming sequence of words passes through morphology as an independent word, resulting in a canonical form and a morphological characterization for each word. They are subject to the normal rules of syntax (which checks on the legitimateness of the combination and checks agreement where applicable) If a sequence of canonical forms and or surface forms is listed as flexible MWE, it is mapped as whole to its translation equivalent in the target language.

In generation, the MWE is introduced in transfer with multiple components, and an indication of the head. It participates in the syntactic rules as if it were a normal combination of these components, and nothing special has to be done.

Semi-flexible MWEs can deal with slightly more complex constructions than fixed MWEs, but MWEs with irregular internal syntactic structure require a treatment in the system similar to the one described for fixed MWEs in the Rosetta system. This can be implemented by treating all MWEs without internal inflection in this way.

Semi-flexible MWEs do not constitute a separate class of MWEs in Rosetta: the only way to deal with them is to treat them as flexible MWEs, to be discussed in the next section.

F.2.5.3 FLEXIBLE MWEs

A principled and very powerful method for dealing with flexible MWEs has been developed in the Rosetta project and implemented in the Rosetta system, for details refer to Rosetta (1994) and especially Schenk (1986), Schenk (1992), and Schenk (1994).

As we have seen above, a flexible MWE can not only occur in a canonical order with contiguous components, it also allows other words to intervene between its components, it allows its component words to occur in different orders, and combinations of permutations and intervention by other words not part of the MWE.

For convenience, we repeat the example given above illustrating this:

[15] a. *Hij heeft gisteren* ***de plaat gepoetst***
b. *Ik dacht dat hij gisteren* ***de plaat*** *wilde* ***poetsen***
c. *Hij* ***poetste de plaat***
d. *Hij* ***poetste*** *gisteren* ***de plaat****.*

In addition, certain flexible MWEs allow for (and require) controlled variation in lexical item choice, e.g. in idiomatic expressions containing bound anaphora such as *to lose one's temper*, where the possessive pronoun varies depending on the subject:

[16] a. *I lost* ***my*** *temper*
b. *You lost* ***your*** *temper*
c. **I lost* ***your*** *temper*
d. **You lost* ***my*** *temper*

Of course, not every flexible MWE allows all of these options, and not all permutations of the components of a flexible MWE are well-formed (e.g. one cannot have **Hij heeft gepoetst plaat de*). The way to account for the properties of flexible MWEs with regard to these phenomena is to assign to a flexible MWE the syntactic structure that it would have as a literal expression: it will then participate in the syntax as a normal expression, and permutations, intrusions by other words or phrases, etc. can occur just as they can occur with these words in their literal interpretation.

Adopting this approach for the Dutch MWE *de plaat poetsen* accounts immediately for (9b), where the verb *poetsen* participates in the formation of verbal clusters in the normal way, just as the expression under its literal interpretation.

The examples (9c,d) are also accounted for by assigning the MWE a normal syntactic structure: it can then be subject to the rule of Verb Second in the normal way, just as under the literal interpretation. It also accounts for the ill-formedness of the example **Hij heeft gepoetst plaat de* given above, since this string is also ill-formed under the literal interpretation.

Flexible MWEs often have restrictions on their syntactic behaviour additional to the ones on normal constructions. Part of these restrictions can be accounted for by the fact that the parts of MWEs often do not have a meaning of their own in the MWE (only the MWE as a whole has a meaning), and that these parts do not refer. These additional restrictions should follow from the design of the grammar and its treatment of idiomatic MWEs, and not be stipulated for each individual MWE.

In the Rosetta system many restrictions on the syntactic behaviour of MWEs or their components are dealt with in a systematic way by the design of the grammar (e.g. restrictions on modification, topicalization, rearrangements in the ”Mittelfeld” in Dutch, pronominalization, etc.).

Other restrictions on MWEs cannot be reduced to general grammatical properties or principles, and must be stipulated as idiosyncratic properties of the MWE. For example, certain expressions can be passivized only under the literal interpretation but not under the idiomatic interpretation.

Furthermore, an MWE must of course be recognized as such and differentiated from its literal counterpart. So at some point in the grammar, the treatment of MWEs must differ from their literal counterparts.

In the approach adopted in Rosetta, this is dealt with as follows. A flexible MWE is described in the lexicon, but it has a number of properties specific to MWEs, in particular

– a syntactic structure

– a list of lexical items making up the MWE

The syntactic structure is not directly represented in the lexicon with the lexical item for the MWE. Instead, a unique name for (reference to) the syntactic structure is specified. This is done because the syntactic structures are quite complex and are shared by multiple MWEs. Using names for syntactic structures rather than the syntactic structures themselves with the lexical items reduces

the effort to add MWEs that require a structure already used for other MWEs, it increases the consistency and makes it easier to maintain the lexicon. The names are called idiom patterns in the Rosetta system The syntactic structures themselves are not S-trees but D-trees. The reasons for this are as follows: S-trees contain nodes with a lot of attribute-value pairs.

It is very difficult to get all the values of the attributes correct by hand (and some manual work is required even if the lexicographer is supported by the system). Second, the nature of the attributes, and especially their values, are rather unstable during development of the system: new attributes are added, existing attributes removed or changed, and especially their values regularly change or are extended during development[29]. D-trees are a much more stable part of the system, and D-trees are, in comparison to S-trees, relatively simple: most nodes have atomic labels for rule names, and only few nodes have attributes (rule parameters).

In analysis, the surface structure of an utterance created by the surface parser and represented in an S-tree, is subject to grammatical rules that check the well-formedness of the S-tree, and gradually modify and reduce this structure to end up with an S-tree in a canonical form (e.g effects of displacements such as Dutch Verb-second, verbal cluster formation, topicalization, etc. are undone), so that the argument structure of the utterance can be checked. Which rules are applied, and to which arguments, is recorded in the syntactic derivation tree for the utterance. At the point in the derivation where syntactic selectional restrictions are checked, it is also checked whether the structure can be analyzed as an MWE. To that end, it is checked whether the structure can be analyzed in accordance with the D-tree of any MWE. If that is the case, the structure is transformed into a simplified structure in which the multiple parts of the MWE have been replaced by a single node, and further only the arguments of the MWE (if any) are present. The resulting structure is then subject to any further rules of the grammar, that, inter alia, check whether the right number of

[29] Though one cannot start working on developing lexical entries for MWEs before a significant part of the syntax for single words has been developed and is reasonably stable, it is also unrealistic to expect that MWEs will only be added to a fully stable and unchanging system.

arguments is present: when this is the case, the full analysis of the utterance containing an MWE yields a derivation tree that is indistinguishable from derivation trees for utterances not containing any MWE.

In generation, the D-tree associated to an skey for an MWE is used to generated the complex structure for the MWE. Once this structure has been created, it is subject to all the normal rules of the grammar, and will participate normally in syntactic processes. The non-referential nature of the MWEs components will automatically account for several restrictions on the MWEs syntactic potential.

F.2 Support Verb Constructions

F.2.1 Linguistic peculiarities of Support Verb Constructions

Linguists define SVCs as a set of verbal constructions, found in many languages, composed by a semantically empty verb (or light/support verb) and a noun (or NP) as its direct or sometimes prepositional object, the whole functioning as a single predicate and, thus, having lexical status. This becomes evident in that the choice of the support verb seems to be semantically arbitrary- i.e. lexically driven- and considered from a multilingual perspective, literal translations of such constructions are hardly ever possible.

"A SVC is the combination of a support verb and a nominal component (possibly introduced by a preposition), where the overall meaning comes from the nominal component. The support verb contributes only general semantic information like tense, aspect, *aktionsart* and causation"(Kuhn 1994), in addition to the all the morpho-syntactic properties needed for the grammaticality of the phrase/sentence.

Light verbs, in formal terms, can be defined as verbs which have an empty or incomplete argument structure (Grimshaw & Mester, 1988).

Generally, the kind of expressions recognized as SVCs are those in which the NP following the support verb has argument structure, i.e. its head is an event nominal, a predicative noun.

[17] give a demonstration, dare dimostrazione;
[18] pay a visit, fare visita;
[19] take into account/consideration, prendere in considerazione.

Especially computational approaches do not generally take into account the kind of similar constructions in which the head of the NP is a non event denoting noun:

[20] take a shower = fare la doccia, take a coffee = prendere un caffè etc. (see. Wierzbicka 1981 -- and Dixon 1991).

Such constructions, though structurally similar, are taken to belong to different areas. While constructions formed with predicative nouns can be treated (at least partially) by means of rules (see. Namer (1998), Khun (1994) for details), those composed by simple/non deverbal nouns appear to be non-compositional, i.e. fully lexically determined.

F.2.1.1 VARIANTS OF THE SUPPORT VERB.

The light verb may have variants carrying different aspectual values: neutral, inchoative, terminative, continuative, causative.

Kuhn' s examples are:

[21] Gewissheit haben, erlangen, verlieren, behalten, geben. ('avere, ? , perdere, mantenere, dare certezza')

Such variants seem to be lexically determined by the support verb- i.e. they are its (quasi-) synonyms, antonyms etc.- but the possible aspectual information they carry is determined by the semantics of the predicative noun.

Support verb variants of the Italian SVCs in *[9]* are *fornire* and *offrire:*

*[22] **Fornisce** certissime dimostrazioni di tutte le cose occulte ('provides certain demonstration od all mysterious (??) things')*

*[23] Recenti risultati **offrono** una seria dimostrazione della validità… ('recent results offer a serious demonstration of the validity...')*

F.2.1.2 Discontinuity.

A second important characteristic of SVC, which makes them particularly hard to describe, is their discontinuity. It is even more a problem as long as SVCs show high variability with respect to it. So, while *give a demonstration (dare dimostrazione)* shows a possible high discontinuity, *take into account (prendere in considerazione)* or *fare visita (pay a visit)* seems to be syntactically more restricted/fixed.

In some cases, internal modification by AdjP or AdvP is free, in others the elements tend preferably to be contiguous, in other cases, instead, modification is blocked: *give a demonstration*, like its Italian counterpart *dare dimostrazione*, is an example of an extremely free SVC, as shown in the following corpus data:

*[24] They were given a **detailed** demonstration of how…*

*[25] Who gave **me** a **faultless** demonstration of captaincy…*

*[26] Ha dato **ampia** dimostrazione ('he has given large/wide demonstration')*

[27] Non posso qui dare dimostrazione *('I cannot give here demonstration').*

On the contrary, *pay a visit* and *fare visita*[30] are examples of a more fixed SVC, though not completely:

- *…a **far visita** a Papa Giovanni*
- *Il magistrato **avrebbe fatto visita** la settimana scorsa a Cossiga*
- *…di una conoscenza cui è venuto a **far visita**…*
- *…cui **farà visita** nell' ultimo turno della stagione*(!)

In the corpus, 68 out of 90 occurrences of the construction, in fact, have the two elements together with no intervening element. We can, the, say that this SVC preferably occur without intervening elements. This observation, nevertheless, cannot be stated as a rule, for cases in which modification is acceptable are still considerable in number.

*[28] Tony gli aveva fatto **quella terribile** visita in ufficio*

*[29] Qualcuno … che faccia **qualche** visita ad Annie in ospedale*

F.2.1.3 Use of the article.

The same kind of variation is observable in the use of articlex. Some constructions do not impose any constraint on the use of the indefinite/definite article, nor on other kinds of determiners, whereas in some constructions the use of one specific article, or no article at all (which is unusual for Italian syntax) is obligatory.

[30] The example of *fare visita* show an inherent problem of MWEs: the degree of idiomaticity. *Fare visita*, with no intervening element is felt to be semantically more cohesive, and more idiomatic than *fare una visita*, though they express the same general sense of the construction, as opposed to another sense of the same construction: the medical one. One could propose here to treat the more fixed, idiomatic construction as a Multi-Word Lexeme, whereas the more free one as a common SVC, but this is still an open topic.

[30] *E dare* ***la*** *dimostrazione che si tratta ('and give the demonstration that it is…')*
[31] *Neil gave* ***the*** *perfect demonstration of…*
[32] *Voleva dare dimostrazione della propria abilità/ del costo sostenuto ('he wanted to give demonstration of his own ability/ of the high expense')*
[33] *Dovevo dare* ***una*** *dimostrazione di buona volontà* ('I had to give a demonstration of good will')
[34] *She gives* ***a*** *demonstration of her skills*

The use of the definite article in *fare visita* seems to be permitted only in very special contexts:

[35] *A fare la solita visita a tua madre;*
[36] *…Feci la mia prima visita negli Stati Uniti*

Conversely, the use of the indefinite article is allowed, although it is less frequent.

Moreover, we observe a correlation between the presence of an article and the possibility of internal modification: the absence of an article seems to block also internal modification; when the article is used, instead, modification is frequent.
The English equivalent is a SVC with a more constrained use of the article: *pay a visit*, whereas in German, along with an equivalent SVC *ein Besuch machen*, there is also a single verb stem: *besuchen*.

We can thus propose, somehow following Kuhn' s approach, a classification of SVCs on the basis of these characteristics:

1- presence of aspectual variants of the support verb[31].

2-free/obligatory/constrained presence of the article

3- degree (grade) of discontinuity/contiguity.

4- type of morpho-syntactic realization: ex. Presence/absence of a preposition, case (for languages that do have cases like German) etc.

F.2.1.4 THE SEMANTICS OF SVCS.

Wierzbicka (1967, 78-81) notices the highly systematic behavior of periphrastic constructions containing the verbs *make, have* and *give*, that seem to be governed by some sort of rules. She argues that such rules are basically semantic in nature, that is they reflect different possible conceptualizations of the same situation.
Only those constructions which have simplex verbs as counterparts are taken into account in her analysis which assumes the predictability of the difference in meaning between the constructions and their simple-verb counterparts. semi-idiomatic expressions like *have a go* are disregarded.

[31] The identification of the possible variants, however, is not yet a clear cut process. Even in literature, we have not found useful criteria for isolating them clearly. The linguist' s intuition is usually employed. Anyway, it must be recognised that up to now it has been paid little attention to the problem of aspectual, and even less to stylistic, variants. For what concerns the Italian language, eventually, we cannot perform an appropriate corpus analysis of stylistic variants, due to the limited dimensions of our corpus.

F.2.1.5 KINDS OF SVCS

Syntactically we can distinguish between 3 types of constructions:

1. NP + HAVE/MAKE/GIVE (aux) + *a* + V-infinitive: ex. *Have a walk* (taken from Wierzbicka)

2. NP$_1$ + HAVE/MAKE/GIVE (aux) + *a* + N: ex. *Have a coffee*.

3. NP$_1$ + HAVE/MAKE/GIVE (aux) + *a* + deverbal N: *have a quarrel*, *make a proposal.*

It is possible to add a further distinction between constructions in which NP$_1$ is the subject of the event denoted by the full verb and those in which it is not, as in *Mary had a visit by her parents.* (NB: this is a passive construction whose active correspondent should be: *the parents paid a visit to Mary*, which still contains a SVC).

F.2.1.5.1 SEMANTIC CHARACTERISTICS

According to the distinctions made above, at the semantic level at least 6 types of SVCs should be recognized.

Not many work have tried to investigate the semantics of such constructions, and anyway, they have considered only few kinds of them. The most informing are Wierzbicka' s (1982) and Dixon' s (1991) investigations on the HAVE *A* and TAKE_A patterns, though they did not considered them SVCs.

F.2.1.5.2 HAVE A V

The main semantic difference between a *have a*_V construction and a simple verb is aspectual.

Ex. *have a walk* vs. *walk.*

This periphrastic construction is an "agentive, experiencer-oriented, atelic, anti-durative and reiterative" (p. 759).

Verbs which cannot be modified by adverbial expressions indicating duration cannot occur in this frame.

The *Have a V* pattern can be subdivided into two semantic subtypes: one where the V is a verb of perception, the other where the V is a verb of bodily action. In the first case the SVC" imply an action ' which could cause one to find out' , the latter imply an action ' which could cause one to feel good' (Wierzbicka p.761). Such differences, it is claimed, are not idiosyncratic, but show some regularities in the language.

F.2.1.5.3 TAKE A V

This pattern seem to imply a deliberate action starting at a precise point and limited in time and apparently involving physical motion. The activity described must be also perceived as unitary by the agent and have a natural beginning and end. This pattern is thus speaker-oriented, conveying his own perceptions and intentions.

"When one takes a walk, one has a definite idea of what one wants to do and how long it is going to take…ex. *take a walk around the pond*" (795). Semi-voluntary actions, as ' yawning' and ' crying' cannot be used with *take.*

F.2.3 COMPUTATIONAL TREATMENT OF SVCs

Among Genelex-inspired frameworks, an interesting study aiming at isolating the defining, or most distinctive, properties –and to formalize them- of different sub-types of "verbal collocations, focusing in particular on restrictions on syntactic and lexical variability in order to allow for a formal representation in a PAROLE-like computational lexicon, has been made within the Danish STO-project -a continuation of the LE-PAROLE morpho-syntactic computational lexicons (Braasch & Olsen 2002).

The constructions under investigation here are verbal bound word combinations spanning, which happens to be not fully compositional, among which find their place als SVCs as defined in this section.

The STO-project does not take into account the semantic level, so that the properties they basically take into consideration are: the syntactic label of the whole construction; the POS of its constituents/ elements. The fact that the various properties highlighted can combine with each other in different ways, giving birth to different subtypes of SVCs, makes it problematic to handle SVCs within a modular description.

To avoid such problems, in the STO-project a method using patterns has been adopted to describe morpho-syntactic features. A pattern is in their sense a generalized description of a particular linguistic behavior consisting of a unique combination of relevant information, expressed in terms of feature-value pairs.

The relevant info the STO-project come up with are:

- whether definiteness and inflection of nouns are free, restricted or impossible: ex. *tage* *<u>kørekort</u>* (lit. ' take driving license') N(Obj.){sing.indef};
- whether passive transformation is allowed or not, and/or what kind of passive construction is permitted[32]: ex. *<u>tage kørekort</u>* VP{no_pass};
- whether insertion of an attributive Adjective Phrase modifying the noun is free, lexically restricted, or not allowed at all: ex. *tage til* *<u>genmæle</u>* (lit. ' take to reply') N{n_a}[33]

Other relevant features they make use of are indications of:

- the continuity of the structure: values= yes,no.
- the stability of the lexical choice of the N: whether variation is not possible, restricted to one single item, restricted to a few items, restricted by semantic type, or restricted to an enumerable lexical set.
- the layer where each type of word combination identified can be described as a lexical unit, is also indicated.

F.2.3.1 SVCS AND LEXICAL FUNCTIONS

SVCs are a good example of purely linguistic restrictions that seem to be best treated by using mel' cukian Lexical Functions. An analysis of LFs and an evaluation of their usefulness in the description of SVCs has been carried out by Balkan & Dirk (1992), in the context of a more wide investigation aiment at the representation of collocations in computational lexicons. More recently, works on LFs within the Meaning Text Model have gone further (cf. Polguère 2000).

Mel' cuk' s LFs are intended to give a systematic lexical description of co-occurrence restrictions of a given lexeme with other lexical items. LFs are different from selection restrictions and cannot be expressed by that mechanism because:

1- LFs express lexical restrictions whereas selection restrictions typically are imposed on the semantics (denotation) of the argument.

2- the order of selection is reversed: i.e. in *commit murder, commit* selectionally restricts its arguments to OFFENCES. At the opposite, the LF analysis would say that *murder* selects *commit* as its support verb.

3- selection restrictions are about World knowledge, that is basically of semantic nature whereas LFs are about language.

[32] NB. In Danish the passive can be made in two ways.

[33] Meaning ' no adjective insertion is possible' ; the other values are *a*, standing for ' restricted adj.' -when insertion is possible but semantically highly restricted. No value is given if modification is possible and only weakly semantically restricted.

Altogether they seem to be a powerful means of describing them but with some limitations: in particular, lexical functions only give a partial semantic specification.

The relevant functions to the analysis and description of SVCs appear to be:

OPER$_n$: indicating a semantically empty verb that takes the noun of the first, second etc. argument of the situation described by the keyword (the noun) as its grammatical subject and the key word as its first object complement.
Ex. OPER$_1$ (*attenzione*): *fare/prestare* 'make/lend' OPER$_1$ (*attention*): *pay* lit. It. 'pagare'.

CAUS/LIQU/: indicating causativity

Ex. CAUS(*attenzione*): *attirare* 'attract'; CAUS(*attention*): *draw*

REAL indicating...:
ex. REAL2(*requirements*): *fulfill/meet/satisfy*.

A problem which has emerged is that it is often possible to have two or more different LFs describing equally well the same expression, each function possibly underlying a different aspect of the construction: for example *make an offer* can be described equally satisfactorily either by the LF LABOR or by LABREAL, or even CAUS. The latter implies some kind of subtle realization or fulfillment of the concept expressed by the keyword.
In conclusion, although LFs seem to be a powerful means of describing SVCs, their application is not straightforward. "There are cases in which the boundaries between lexical functions need to be defined more tightly".

F.2.3.2 POSSIBLE DIAGNOSTICS TO IDENTIFY SVCS, AND TO ESTABLISH THEIR DEGREE OF DISCONTINUITY.

F.2.3.2.1 INSERTION

Serves to test whether "foreign" material may be inserted into parts of the unit, to make internal modification of parts of the unit.

*[37] Take a decision → take an **important** decision*

But

*[38] take place → * take **good** place*

This test of course does not involve the obligatoriness/ optionality of arguments, which do not affect the parts of the unit (*take into account the problem/ take the problem into account*); this might be regarded instead as a syntactic problem, concerning the possibilities of realization positions of arguments.

F.2.3.2.2 EXTRACTION

This serves to show if a component may be displaced w.r.t. it usual position in the expression. 4 are the possible tests to run: passivization, clefting, left or right dislocation, deletion.

[39] Pay attention to → attention was paid to...

But

*[40] Pay attention to→*it was attention which was paid to...*

Comments: here it is not clear if all these 4 tests should be applicable to the same expression in order for it to be a collocation, or just one will suffice.

F.2.3.2.3 PROFORMATION

It the replacement of a constituent in the lexical unit by a proform/pronoun.

[41] Take a decision → take it again

But

*[42] Take place → * take it again*
*[43] Pay attention to → * pay some again.*

F.2.4 Possible levels of analysis

In the following we will suggest a possible layered structuring for the description of complex nominals and support verb constructions. Expressions can be described/represented at one ore more levels, depending on the desired richness of information. Here, the relevant types of information and constraints that have been judged to be needed are not necessarily presented in a formalized way.
We, thus, propose five levels of analysis, each adding different types of information of increasing complexity and depth. These are respectively: the "sequence of lexemes" level[34], the morphological level, the syntactic level, the semantic level and the multilingual level.
The examples shown in the following pages have been extracted from two corpora, namely the PAROLE Corpus for Italian and the British National Corpus for English (Burnard 2001).

F.2.4.1- MWEs AS SEQUENCES OF LEXEMES

At this level, MWEs are treated just as a sequence of lexemes, thus allowing its elements just to inflect. For each MWUnit we can specify the syntactic category of the whole, frequency in the corpus, register, domain, corresponding single verbs for SVC, hyponym and/or synonym for CNs and other lexical relations. Moreover, esp. for CNs a reference to their semantic class and/or a SIMPLE-like template type appears to be useful.
Practically, the MW-Unit would be considered like a single lexical item, except for the possibility of inflection for one or both formatives (usually it is necessary just for the syntactic head). This, however, is a very elementary level of description, giving no information about the internal constitution of MWEs, and not at all economic for the building of a lexicon.

Ex.
1

<fare visita> = MU= *fare* + MU= *visita*
POS: V
(**Complement**: indirect object, realised as a PP_a: *fare visita a qualcuno*)
Frequency in Parole: 68 as unit, 90 both as a unit and with intervening elements.
Domain: general
Related lexical verb: *visitare*1

<pay a visit> = MU= *pay* + MU= *a* + MU= *visit*
POS: V
(**Complement**: indirect object, realized as a PP_to: *pay a visit to sb*)
Frequency in BNC:
Domain: general

[34] NB: CNs could simplistically be treated as words- with- spaces, but such an approach is not appropriate for SVCs because the verb must be allowed to inflect, and it would cost too much in terms of storage to encode all possible forms of the verb as alternative realisations or even as to create so many entries for the same MWE as are the forms of the verb. Such an approach, though not satisfying, is viable for CNs, for which at least we have to enter two alternatives: one for the singular head noun and another in case the head noun is plural.

Register: (?)
Related lexical verb: *to visit*_1

2

<fucile da caccia>

POS: N
Frequency: 9
Domain: general/ sport
Register: (?)
Hyperonym: SemU= fucile
Semantic class/template type: INSTRUMENTS

<hunting rifle> = MU= *hunt(ing)_x* + MU= *rifle*
POS: N
Frequency in BNC: 3
Domain: general/ hunting
Register: (?)
Hyperonym: *rifle*

This description will account for expressions that can or happen not to have internal modification. This type of encoding would probably suffice for recognition and interpretation tasks (still internal modification is not considered) since it does not matter to constraint the noun modifier of CNs and the nominal of SVC to be plural or singular etc. We can suppose they occur correctly in the texts treated. As for generation, instead, the level is definitely not sufficient.

F.2.4.2 2- MORPHOLOGICAL LEVEL

The second level of description proposed, the morphological level, should allow the lexicographer to represent constraints and restrictions on the morphological form of the modifier, and of the preposition in CNs (usually the head is totally free).

<fare visita>
MU= fare + (MU)= visita (restriction: no plural)

<fucile da caccia>
Mu = fucile + Mu =da(X) + Mu= caccia (no plur)

Here we can specify for SVCs the possibility of the nominal to occur in the plural form, or the constraints on it: in *fare visita* as in *pay a visit* the noun must obligatorily occur in the singular. In a parallel way, in CNs the form of the modifier is normally constrained.
Therefore, at this level two possible strategies of codification are possible:

a) express the link to the MU for the head and leave the modifier, and the nominal as no better specified forms.

b) express the link to both MU and contemporarily express the restrictions on the appearance/ form of the modifier/ predicative nominal.

We prefer the second approach.

F.2.4.3 3- SYNTACTIC LEVEL

A third level of analysis would be one at which one can describe the internal and external syntactic composition of MWEs. For each constructions we express the links to the syntactic units it is formed with and indications about the head of the Phrase.

At this level, it is important also to make explicit the external complementation pattern of the whole construction. This kind of information is particularly useful for SVCs, but appears to be interesting for CNs as well.

The complementation pattern of the whole seems to be inherited from the nominal element: i.e. from the semantic head.

At this stage, moreover, it will necessary to specify the possibility for internal modification: relevant information here will be if modification is blocked, allowed but restricted, or totally free. In this last case, we could, additionally, think of indicating the preferred type of modifier (adjP, advbP etc).

At the syntactic level, moreover, the restrictions applying in the use of articles, if any, have to be specified.

Finally, it is desirable to have the possibility of expressing the syntactic positions realized by the arguments/roles to be expressed in the semantic layer, if present.

<fare visita> ‘pay a visit’

Syntactic composition: synU= *fare* + synU= *visita*
Syntactic head = V= *fare*
Passivisation: yes
Complement extraction: yes
Use of articles: no article or indefinite = allowed; definite = No.
Syntactic positions(w.r.t. *fare,* the syntactic head Pos0= Subj, pos2= IndObj (PP_a).
Modification:
Premodification and/or postmodification of the nominal is possible.
Premodification by AdjPh: restricted to the use of the indefinite article;
Postmodification of the verb by AdvP: freely allowed.
Postmodification: free, adjP preferred.

<fucile da caccia> ‘hunting rifle’

Syntactic composition: synU= fucile + PP= da(P) caccia(N)
Syntactic head = N =*fucile*
Modifier: PP= *da caccia*
Use of article: blocked for the modifier.
Complement Structure: NO
Modification:
AdjP Pre-modification of the head: allowed.
Post-modification of the head: allowed.

F.2.4.4- SEMANTIC LEVEL

At this layer, for SVCs we need to express the fact that the verb is (nearly) semantically empty, and, thus, to indicate the predicative nominal as the semantic head of the construction. It is, nonetheless, be useful to express the possible variants of the support verb, with their aspectual values, for which Lexical Functions à la Mel’cuk could be used. Moreover, we need to give indications about the semantic/logical arguments of the construction, inherited from the event noun, and, additionally, link them to the syntactic positions described in the syntactic layer.

For CNs, instead, it is at the semantic level that one must express the implicit semantic relation that links the formatives, and the idiosyncratic bit of meaning one must add to obtain the correct full interpretation, where relevant. In our examples we have exploited the Qualia Structure model, as it is presented in Pustejovsky (1995) and implemented in the SIMPLE lexicons.
In order to describe the semantic internal structure of CNs, however, it is worth noticing here that other descriptive frameworks exist that are able to capture different semantic component of lexical items and express the relationships between items. Among this we mention Frame-semantics (Fillmore et al. 2001), the NOMLEX project (MacLeod et al. 1998).

<fare visita>

semantic head: SemU= visita
Supp Vb: fare (aspect: causative)
Variants SuppVb: No
Argument Structure (of *visita*) **Arg0**= agent
Arg1= beneficiary

<fucile da caccia>

semantic head: SemU= fucile (rifle)
modifier: SemU= caccia (the hunting)
implicit semantic relation: telic:used for: *cacciare* (to hunt);
used by: *umani/ cacciatore* (humans/hunters);
idiosyncratic meaning: is used to shoot animals, for food or as a sport.
used by: *umani/ cacciatore* (humans/hunters)

alternatively, specifying the semantics of the event:

used for (activity): *uccidere* (to kill)
object of the activity: *animali* (*selvatici*) ((wild) animals)
used by: *umani/ cacciatore* (humans/hunters)

F.2.4.3 5- MULTILINGUAL LEVEL

Here we can establish the links between the equivalents in different languagex..

<fare visita>

translation equivalents : ingl <pay a visit>
fr. <faire une visite>,
Dtsc. <besuchen>, <ein Besuch machen>

<fucile da caccia>

translation equivalents: ingl. Lex= NC <hunting knife>
fr. Lex= NC <fusil de chasse>
Dt. Lex= NC <Jagdgewehr>

REFERENCES

- Balkan, L. and D. Heylen. 1993 *Collocations and the lexicalization of semantic operations,* Technical Report ET-10/75.

- Balkan, Lorna. et al 1992 "Analysis of Lexical Functions" *Characterising Collocations,.*

- Boucher, Paul et al. 1993 "Compounds: An Intelligent Tutoring System for Learning to Use Compounds in English." Interim Report n. 718. IRISA. April.

- Braasch, A. and S.Olsen, 2002. "Towards a Strategy for a Representation of Collocations - Extending the Danish PAROLE-lexicon". in Proceedings of the 2nd International Conference on Language Resources & Evaluation, Greece.

- Braasch, A. and S.Olsen, 2002. "Formalised Representation of Collocations in a Danish Computational Lexicon" in U. Heid & al., (eds.) Proceedings of the Ninth EURALEX Congress. Stuttgart. p.475-488.

- Breidt, Elisabeth. 1994 *Terminology for Bilingual Dictionaries in Computational Lexicography.* Seminar fuer Sprachwissenschaft. Tuebingen: Universitaet Tuebingen.

- Busa, Federica and Micheal Johnston. 1996 "Cross-Linguistic Semantics for Complex Nominals in the Generative Lexicons." Proceedings of the *AISB Workshop on Multilinguality in the Lexicon.* Brighton UK: University of Sussex, April.

- Calzolari Nicoletta, Lenci Alessandro, Quochi Valeria, "Towards Multiword and Multilingual Lexicons. Between Theory and Practice". *Proceedings LP2002,* September 2-6, 2002, Japan: Meikai University.

- Dixon, Robert M. W. 1991. *A New Approach to English Grammar, on Semantic Principlex.* Oxford, New York : OUP, Ch. 11, 336-354.

- Downing, Pamela. 1977 "On the creation and Use of English Compound Nouns."Language 53: 810-842.

- Fabre, Cécile and Pascal Sébillot. 1995 "Calculability of the semantics of English Nominal Compounds: Combining General Linguistic Rules and Corpus-based Semantic Information." INRIA: Interim Report n.2742 (programme 3).

- Grimshaw, Jane and Armin Mester. 1988 "Light Verbs and θ-Marking." *Linguistic Inquiry*: 19, 205-232

- ten Hacken P., D. Maas and B. Maegaard. "Dictionaries in Eurotra." *Studies in MT and NLP* Vol.1.

- Johnston, Micheal and Federica Busa. 1999 "Qualia structures and the compositional interpretation of compound." in E.Viegas(ed.) *Breadth and depth of semantic lexicons.* Dordrecht: Kluwer Academic Publishers.

- Khun, Jonas 1994 "The treatment of support verb constructions in HPSG-based Machine Translation – a summary." Institut fuer Machinelle Sprachverarbeitung, Universitaet Stuttgart.

- Leonard, Rosemary. 1984 *The Interpretation of English Noun Sequences on the Computer*. Amsterdam: Elsevier Science Publishers,.

- Levi, Judith N. 1978 *The syntax and semantics of Complex nominals*. New York: Accademic Press,.

- Lieber, Rochelle. 1983 "Argument Linking and Compounds in English."Linguistic Inquiry 14, 251-283.

- Lyons, John. 1977 "The Lexicon." Semantics. Cambridge: CUP,. 512-569.

- Namer, F. 1998 "Support Verb Constructions" in F. van Eynde and P.Schmidt (eds.) *Linguistic specifications for typed feature structure formalisms*. Studies in MT and NLP Vol. 10. Brussels: CEC,. 315-343.

- Pennanen, Esko. 1980 "On the Function and Behaviour of Stress in English Noun Compounds." English Studies 61, 252-263.

- Petrocelli, Simonetta. 1992 "La Composizione Nominale in Italiano e in Tedesco."Studi di Linguistica Italiana Teorica e Applicata. XXI, 65-82.

- Polguére, Alan. 2000. "A 'Natural" Lexicalization Model for Language Generation." in *Proceedings of the Fourth Simposium on NLP (SNLP2000)*. Chiangmai, Thailand, 10-12 May 2000, pp.37-50.

- Quirk, Randolph et al. 1985 *A Comprehensive Grammar of the English Language*.New York: Longman,.

- Segond, Frédérique and Elisabeth Breidt. 1992 "IDAREX: Formal Description of German and French Multi-Word Expressions with Finite State Technology."Interim Report. nn, November.

- Warren, Beatrice, ?1978, "Semantic Pattern of Noun-Noun Compounds."Acta Gothenburgensis Studies in English.

- Wierzbicka, Anna. 1981 "Why can you *have a drink* when you can' t *have an eat*?" *Language* 58 (1982), 753-799.

Appendix G: EAGLES/ISLE Enlargement to Asian Languages

Introductory remarks

A crucial aspect in establishing a real and broad consensus is played by communication and sharing of information among many groups active in the field. For this reason we involved also Asian collegues in the ISLE initiative, and we are exploring ways of establishing formal links with them.

An enlargement of the group to involve also Asian languages has been pursued and representatives of Chinese, Japanese, Korean, and Thai and other Asian languages have contributed to ISLE work and participated in some ISLE workshops. Also the newly formed *Asian Federation of Natural Language Processing Associations* (AFNLPA), chaired by J. Tsujii, has declared interest in the ISLE standardization initiative, and the gradual involvement of Asian groups in the ISLE initiative is being now strengthened.

The contribution from Asian partners consists of two sections. The first one deals with general issues concerning the lexicon, its internal organization, the definition of its atoms, i.e. what a lexical entry is, in particular, from the point of view of Asian language typology.

The second one deals with a more applicative point of view, reporting on the results of the application of the EAGLES recommendations for morphosyntax to a set of Asian language.

Issues in the Structure of Lexicons: A Multilingual Perspective

Defining Lexical Entry / Lemma

A lexicon consists of lexical entries (i.e. the atoms) organized in a specific lexical structure. Standard and sharable multilingual lexical resources must have a cross-linguistically robust definition of the atoms and the lexical structure.

The definition of a lexical atom determines the content of the lexicon. Given the wide typological variations of the world's languages, a definition of the lexical entry must be robust and versatile such that it can apply as effectively and naturally to all languages.

We assume that the standard lexical entries are lemmas of a particular language. And the set of lemmas in a language is the result of optimal lemmatization. In other words, the set of lexical entries in any language is the optimal way to itemize the language.

We further hypothesize that the optimization is obtained by the following definition: lemmas are conceptual atoms which are also morpho-syntactically autonomous. In any language, there may be conceptual atoms which are not morpho-syntactically autonomous, and vice versa. Hence, it is this simultaneous requirement which ensures that the lemmas are the optimal way to itemize the grammatical knowledge of a language.

In consequence, lemma will include all the following linguistic units: Words, Stems, Affixes, and Clitics. Words are the traditional units adopted by lexicographer. However, words as lemmas may not be tenable in languages with rich morpho-phonology, such as the Austronesian languages. In these languages, a word is always composed of a stem and several affixes, and the phonological form of the word is dependent on the morpho-phonological rules involved. In other words, a word has too many variations to list, and all the variations can be predicted by the rule-governed combination of stems and affixes. In these languages, the optimal lemmas will consist of stems but not words. Similarly, affixes and clitics are conceptual atoms that have autonomous morphological slots (even though often phonologically dependent on their hosts.)

A consequence of such a view to lexicon and lemma is that Lexical Idiosyncrasies occur at the lemma level. Since lexical rules apply lemmas in each language, this is also the level where exceptions will be found.

In sum, we suggest that a definition of lemma must be specified in the metadata of a language lexicon. It must state whether it has word-based lemmas or stem-based lemmas. The lexicon must also allow for morpho-lexical rules to be encoded as lemmas. These will include inflectional/derivational morphology, cliticization, as well as compounding. These morpho-lexical units must be categorized according to the input and output of their rule application.

Last, for languages whose lemmas (such as words) are not conventionally demarcated, it is essential that a segmentation as a standard process in defining lemmas is defined/stipulated for that language.

Orthography and the Implicit Structure of the Lexicon

Each lexicon has an implicit internal structure, which is often stipulated by orthographic conventions, such as alphabetical in various European languages. Roman alphabets differ from Slavic alphabets. And languages using the same alphabet sets may have lesser variations among themselves. Non-alphabetic languages follow different conventions, such as the radical-stroke system of Chinese. Even though content of electronic lexicons can be randomly accessed without following the conventional order, users do perceive lexicons as sequential databases and access electronic lexicons based on that assumption. In addition, default inheritance mechanisms often assume such implicit structure. In sum, any lexicon must be clearly marked for its orthography as well as its conventionalized orthographic order.

It must also be noted that one language can conventionalizes more than one set of orthographies. The best-known example is perhaps Japanese with its three orthographies: Kanji, Katakana, and Hiragana. Note that there is a set of rules in Japanese, stipulating when to use which convention.

It is also important to remark that with the rising popularity of the web and the easier access of multilingual information, a lot of English words associated with new technologies or new products have found their way into other languages. What is interesting is that these terms maintain may their original orthography and (to a large extent) pronunciation. The words adopted in Chinese include IBM, and ADSL.

Once an orthographic convention is established in a language to represent loan words, the convention can often be adopted to transcribe other loan words, regardless of whether these loan words come from the same convention. Japanese Hiragana is highly conventionalized for this purpose. In Mandarin Chinese, the English alphabets are now used to transcribe all loan words. For example, a loan from Taiwanese lau-ko-ko ‘to be old and senile’ is written as lkk.

The use of orthography is made even more complicated in a few scores of words that are themselves Code-Mixers. Note that all the code-mixed words are pronounced according to its orthography. That is the Chinese character part pronounced according to Chinese and the English alphabet part pronounced according to (an imperfect) English phonology. The ‘Q’ sound, for instance, is not phonemic in Chinese.

a-Q a typical Chinese who is cynical and fatalistic, from a famous novel
K-shu1 to hit the book
C-zhao4bei1 C-cup (as in a bra)
A-qian2 to ill-gain money, gained but not earned
ah-sir (Hong Kong Cantonese) a police officer

Since these words are bona fide lexical entries in the language, they offer several challenges to a standard lexicon structure. Lexical Information will be lost if orthographic information is lost.

The solution we suggest is that orthographic information be encoded in a standard lexicon. At lexicon level, orthographic conventions of the target language must be stipulated. In addition to set of orthographic atoms, additional information may include the lexicon structure conventionalized by the orthography (alphabetical order – with alphabet sets identified, radical classification etc.)

This information should reflect the representation adopted in the lexicon (e.g. Pinyin Romanization for Chinese). Concerning how word/lemma boundaries are marked by the orthography, at entry level, orthographic convention will be marked on each entry, including the possibility of code-mixed orthography. Of course, for the majority of lemmas in the majority of languages, unmarked default will be the dominant orthography stipulated for the whole lexicon.

Directionality in Multilingual Lexicon

A non-trivial issue regarding multi-lingual lexicon is whether it is simply composed of linked mono-lingual lexicons or if it is actually a collection of multi-lingual records. We show by the following examples that change of direction may affect cross-lingual lexical correspondences. Hence, the linked mono-lingual lexicon is too simplistic a model for multilingual lexicon.

English to Chinese

Phoenix
Feng4huang2
➔***A bird in Egyptian mythology that lived in the desert for 500 years and then consumed itself by fire, later to rise renewed from its ashes.***

Chinese to English
Feng4huang2
Phoenix
➔***A bird in Chinese mythology that always showed up in a pair: the male feng4 and the female huang2. They symbolized love and marital bliss.***

In the above example, we see that even though phoenix-Feng4huang2 seem to be a one-to-one pair, the directionality of the bilingual lexicon actually changed the interpretation. In other words, although they are conventionalized to translate each other, phoenix and Feng4huang2 in their own source languages have very different meanings.

The following example is even more straightforward. It is a case where there is a many-to-one mapping relation between Chinese and English.

Chinese to English
bo2bo5
uncle
An elder brother one's father
shu2shu5
uncle
A younger brother of one's father

jiu4jiu5
uncle
A brother of one’s mother

English to Chinese
uncle
bo2bo5 or shu2shu5, or jiu4jiu5

A possible solution to solve this issue is that directionality be marked in multilingual lexicon. A possible place where to stipulate such information is in the metadata. One could adopt OLACMS and stipulate that

Subject.language: the language being described
Language: the language used in description
In an English-to-Chinese lexicon, English will be the Subject.Language, and Chinese will be the Language.
In addition, within the lexicon, we must allow categorical mismatches between Language and Subject.language. Hence, one must be able to specify categories of both Language and Subject.language.

Feedback to the EAGLES proposal from Asian languages

This section describes some feedback to the EAGLES proposal on morphosyntactic framework. The target languages of analysis are Chinese, Formosan, Hindi, Japanese, Korean, Thai.
To obtain the feedback from each language, we first distributed the EAGLES document on morphosyntax (EAGLES, 1996) and asked each contributor to check to what extent the EAGLES framework was applicable to his/her language. This work was done by comparing the EAGLES framework with linguistic phenomena appearing in existing Asian language resources such as lexicon and corpora. The results were then collected and compiled into a table shown in table-G1[35]. As a starting point, only nouns and verbs were analyzed. In the following, we discuss applicability of each feature proposed in the EAGLES documents, and summarize to propose extensions for Asian languages.

1 NOUN

The EAGLES proposal defines Type (L1), Gender (L1), Number (L1), Case (L2) and Countability (L2a) for nouns.

(1) Type
Noun type "common noun" and "proper noun" are applicable to all languages. And all except Hindi need a new noun type "classifier". Generally, languages which do not distinguish singularity and plurality of noun tend to use classifier to denote the number of the object. And the classifier is determined based on the semantic type of objects. The following are examples of a classifier; each of which is a translation of "two dogs".

Chinese:	*liang3*	*zhi1*	*gou3*
	two	CLS	dog
Japanese:	*inu*	*ni*	*hiki*
	dog	two	CLS
or			
	ni	*hiki-no*	*inu*
	two	CLS-GEN	dog
Korean:	*kai*	*twu*	*mali*
	dog	two	CLS
Thai:	*mha*	*song*	*tua*
	dog	two	CLS

Korean needs additional type "dependent" which is always used in accompanying with another noun.
There is no description of compound nouns in the EAGLES proposal. It would be controversial in creating a new type "compound nouns" as a subclass of noun. Further investigation would be necessary on this issue.

(2) Case
Case is not defined as a feature of nouns in all languages. It is marked in different way. For example, in Hindi, Japanese and Korean, case is marked by particles (postpositions) as follows:

Hindi:	Mary *ne*	John *ko*	kiss kiya
Japanese:	Mary *g*a	John *ni*	kisu sita

[35] Since we do not have enough feedback from Formosan, it is missing in the table.

Korean:	Mary *ga*	John *eykey*	khisu hayssata
	Mary-NOM	John-ACC	kissed

Case is marked by position and meaning of nouns in Thai. There is no explicit case marking in Chinese. Thematic/argument role marking is prefered rather than case in Chinese.

(3) Gender, Number, Countability
These features are not applicable to all languages except Hindi. Hindi needs additional feature Animacy (animate/inanimate), which is more semantic oriented. In Japanese and Korean, affixes indicating plurality are used, but it can attach to the limited class of nouns.

2 VERB

The EAGLES proposal defines Type, Finiteness, Verb form/Mood, Tense, Person, Number and Gender as level 1 features of verbs.

(1) Type
Verb type "main", "auxiliary" and "copulative" are applicable to all languages. Additional type "support verb" is necessary for Hindi, Japanese and Korean. Support verbs derive denominal verbs, for example;

Korean:	kongpu	*hata*
Japanese:	benkyo	*suru*
	study	do.

Similar to nouns, investigation on compound verbs would be necessary. In particular Hindi needs "compound verb" as a separate type.
In Chinese and Thai, distinction between state verbs and action verbs is very important. This information should be described in a lexicon.
Modal is expressed in various way, such as auxiliaries, particles and suffixes.

(2) Finiteness
This features is not applicable to all languages.

(3) Verb form
Since Chinese and Thai are isolating languages, verbs do not conjugate. Japanese and Korean verbs conjugate, but their form is determined by the succeeding word instead of finiteness, mood and so on.

(4) Mood
Mood is expressed by auxiliaries and particles in Japanese and Thai, and by suffixes in Korean.

(5) Tense
Tense is expressed by auxiliaries, suffixes and adverbs rather than verb conjugation.

(6) Person, Number, Gender
These features are not applicable to all languages except Hindi.

(7) Aspect

Aspect is expressed by auxiliaries, suffixes and adverbs (Chinese). Aspect is closely related to tense. A representation framework for tense and aspect is an open question.

(8) Voice
Feature value active/passive are applicable to all language.

(9) Reflexivity
Reflexivity is not applicable to all languages.

3 Summary

There is clear contrast between inflectional language (Hindi) and others. But interesting thing is that Hindi shares some features with other Asian languages. For example, since other Asian languages are not inflectional, information marked in terms of inflection in many European languages is expressed by other means such as auxiliaries, particles and affixes. Hindi also express this kind of information in the same way as other Asian languages even though it is an inflectional language. It is interesting to introduce linguistic typological viewpoint in analysis.

In the EAGLES framework, the classification of information seems to be influenced too much by surface representation (inflection). For example, we could have consensus on attribute values of case, but how it is realized depends on languages. They are represented in terms of other means, such as postpositions (Hindi, Japanese and Korean). It would be better to distinguish between information to be described in a lexicon and how it is expressed in surface representation. In comparison table (table-G1), we try to distinguish these; information to be described in a lexicon is placed in the colored line, and its representation means in just below it (non-colored line).

4 Proposals

The following is some proposals to the EAGLES framework from Asian languages analyzed in this work. Note that this is far from complete list. Because we have analyzed only nouns and verbs in six languages. We hope this work could be the first step to make the EAGLES framework be the international standard.

(1) New class "classifier" should be created as a subclass of noun at level 1.

(2) Affixes which is currently classified in Residual plays important role in many Asian languages. In particular, affixes play a crucial role in Formosan. Therefore it should be promoted to the first class category.

(3) Classification of adposition should be more precise for agglutinative languages (Japanese and Korean). This precise classification could be level 2b.

(4) Classification of honorific system would be necessary at level 2b.

	EAGLES	**Chinese**	**Hindi**	**Japanese**	**Korean**	**Thai**
		(isolating)	(inflectional)	(agglutinative)	(agglutinative)	(isolating)
NOUN						
Type	common/proper	< +classifier	<	< +classifier	< +classifer,dpendent	< +classifier
Gender	m/f/n	N/A	m/f	N/A	N/A	N/A
			suffix			
Number	sg/pl	N/A	sg/pl	N/A	N/A	N/A
			suffix	(affix)	(affix)	
Case	nom/acc/gen/dat	N/A	dir/oblique	N/A	N/A	N/A
			postposition	postposition	postposition	position, meaning
Countability	count/mass	N/A	count/mass	N/A	N/A	N/A
			?			
VERB						
Type	main/aux/mod/cop/s-aux	main/mod/cop	main/aux	main/aux/cop/support	main/aux/support	main(state/action)/aux/cop
Finiteness	fin/nofin	N/A	N/A	N/A	N/A	N/A
Verb Form		N/A	N/A	N/A	N/A	N/A
				conjugation	conjugation	
Mood		N/A	?	concl/supp/expc/impr/intr/…	ind/imp/intr/excl/sug	?
			aux	aux, particle	suffix	aux, particle
Tense	pres/fut/past	N/A	<	pres/past	<	<
			aux, suffix	conj, aux	affix	aux, adv
Person/Number/Gender		N/A	<	N/A	N/A	N/A
			infl			

Aspect		perf/prog/exp	perf/cont/habit	perf/prog/stat/exp/fut/…	perf/prog	start/state/finish
		aux, adv	aux, suffix	aux	affix, aux	aux
Voice	active/passive	< +disposal	<	<	<	<
		?	?	aux	affix, aux	aux, pron, pp
Reflexivity		N/A	?	N/A	N/A	N/A

Legend
N/A: not applicable
<: same as EAGLES
?: not investigated thoroughy

Table G1

Appendix H: Resources for spoken language and multimodal lexica in multilingual contexts

H.1 Introduction

H.1.1 General objectives of this report

The ISLE goal of creating consensus based resources includes integrated fields such as speech-to-peech translation. In this area, the Spoken Language Technologies of Automatic Speech Recognition (ASR) and Speech Synthesis cooperate via well-defined interfaces with the written language technologies based on Natural Language Processing, Computational Lexicography and related disciplines. However, Spoken Language Lexicography goes far beyond this area. The main objective of this report is to delimit the field of Spoken Language Lexicography, to show the similarities and differences between this field and that of text-oriented lexicography, and to provide pointers to integrating current work on Spoken Language Lexicography with results of the ISLE project.

H.1.2 Procedural and declarative dimensions of Spoken Language Lexicography

For many reasons, which will be outlined in this report, the area of Spoken Language Lexicography is largely complementary to that of text-oriented lexicography as it is currently understood, both in procedural and in declarative respects. The procedural dimensions of Spoken Language Lexicography - lexical acquisition and lexical access - differ greatly from these dimensions in text-oriented lexicography. There are two main areas of difference. First, spoken language lexica have a wide variety of users operating with a wide variety of application specifications, which will be summarised in the body of this report. The use profile ranges from from consultation of pronunciation information in conventional dictionaries to complex pronunciation databases for speech recognition and synthesis systems.

Second, spoken language lexicography is ultimately based on spoken language corpora, i.e. not simply on transcriptions (which in general provide restricted normative pronunciation representations)but on transcriptions and annotations of actual recordings. A further complexity arises from the increasing acceptance of the fact that speech is multimodal, and therefore information from parallel visual and acoustic signal streams needs to be included in spoken language corpora and spoken language lexica.

The main declarative distinctions between Spoken Language Lexicography and text-orientedlexicography are summarised in basic semiotic terms as follows. Onomasiological lexical information for spoken language includes pronunciation variation, lexical prosody, the interfaces between the lexicon and pronunication detail, the lexicon and sen-tencephonology and discourse, fixed prosody in multi-word expressions, and lexicalised clitic constructions and other functional units, none of which figure in conventional lexicography. In view of the complexity of pronunciation information, compounded by information from parallel prosodic and multimodal data streams, the notion of a *Surface Uni*t, *Surf*U, will need to be defined in future work.

Semasiological lexical information for spoken language includes providing for *light noun*s, as well as *light verb*s, which are used in *ad hoc* nonce formations (where precision is sacrificed to fast lexical recall), lexicalised hesitation phenomena or so-called pause fillers (which actually have distinct discourse structuring meanings and vary from language to language), speech act verbs, subjective adverbs, and a wide variety of discourse particles.

The subdomain of pragmatic lexical information, for which a notion of *Pragmatic Uni*t, *Prag*U, will have to be provided in future work, is much more prominent than in text-oriented lexicography; it pertains to discourse functionality, style level, and register or sub-language appropriateness.

H.1.3 Scope of this report

The present report cannot deal with all the aspects of Spoken Language Lexicography which have been outlined here. The report will concentrate on those aspects of Spoken Language Lexicography which are most closely relatable to the ISLE *Basic Notion*s, with particular reference to the MILE *(Multilingual ISLE Lexical Entr*y). Some of the concepts discussed in this report have been reported on in presentations as the METALEX approach to lexical resource standardization.

The report builds on the SAM, EAGLES and ISLE standardization work funded by the European Commission, and provides a systematic framework for the specification and design of multimodal multilingual lexica. However, it does not exemplify such lexica or such lexical entries; instantiation of such lexica from each of the languages of the European Union (and typologically representative languages outside the EU) is difficult and and extensive task which will need to be performed at a later date and in another context. The report does not propose a definition of metadata for describing spoken language lexica.

The issue of lexical metadata has been surveyed in the ISLE Metadata Initiative (IMDI), using the ISLE bottom-up approach. However, IMDI has not been able to address explicitly the very different field of metadata for Spoken Language Lexicography. Since this field clearly differs in many decisive ways from that of text-oriented lexicography it will need to be treated at a later date and in a different context, taking into account the distinctions described in the present report.

Recommendations for Spoken Language Lexica as required in the applications-oriented Spoken Language Technologies were developed and published in the EAGLES Phase I Spoken Language Working Group (SLWG) [11], and recommendations for Spoken Language terminology lexica were developed and published by the EAGLES Phase II Spoken Language Working Group (SLWG) [10]. These publications constitute the starting point for the present report. In view of the complexities of the area and rapid developments in the integration of acoustic and multimodal components of speech, this report provides a survey and an analysis of the field; specific recommendations which go beyond previous EAGLES SLWG recommendations would be premature.

H.1.4 Background considerations

The main contexts in which spoken language and multimodal lexica are relevant for multilingual contexts are in localisable speech technology systems (in automatic speech recognition and speech synthesis), in speech-to-speech translation, and in other less striking contexts such as the provision of pronunciation information for each language in human readable bilingual or multilingual dictionaries.

The difficult part of multilingual lexicon development for spoken language lies in the coordination of the corpus vocabularies for the languages concerned. First, spoken language system development uses relatively small corpora of transcriptions (perhaps up to several hundred thousand words, yielding a lexicon of several tens of thousands of words, depending on application type). These corpora are expensive and very labour-intensive to make, with real-time factors of between 50 and 500 to transcribe and annotate, again depending on specification. A mere hour of speech, say 20 pages of relatively close-typed transcription, would therefore take between about 1 and 12 weeks to process as a resource for lexicon acquisition, depending on the lexicon requirements specification. Second, in a speech-to-speech translation lexicon the scenario constrained corpus lexicon requirement which invariably has to be met for spoken language system lexica is overstrained by the need to translate from a given corpus lexicon into a translation-generated lexicon in the target language which is by definition not a corpus lexicon, and to process these lexical entries in the target language. The need to process items which are not in the corpus lexicon but need to be accounted for quasi-compositionally is known as the *Out Of Vocabulary* (OOV) problem; this problem is compounded by the translation situation. The path towards a solution to these problems is more general than these specific examples might suggest. Consequently, the present contribution presents a preliminary clarification and systematisation of resources for spoken language lexica with a view to developing standards and resources in this area, and builds on a number of previously collated sources of information. The basic sources are [11], [14].

H.1.5 Spoken language lexicography

A central problem in discussing resources for spoken language lexica emerges from the fact that there is no unified notion of lexicography for spoken language, and therefore no relatively homogeneous guild of lexicographers as there is for written language. Many disciplines, independently of each other, manufacture spoken language lexica. The reason for this lexicographic inhomogeneity lies in the wide range of uses for which lexical information on spoken language is required, some of which are listed here:
General lexica:
– transcription of pronunciation information as a data category in written language lexica,
– pronunciation lexica (orthographic wordlists with phonemic transcriptions),
– rhyming lexica,
– wordlists, glossaries, and lexica for unwritten languages;

Machine lexica for human use:
– transcription and audio output for pronunciation in hyperlexica,
– audio and video concordances (wordlists, pre-compiled or generated on-the-fly)
mapped to timestamps in audio and video recording files);

Machine lexica for written language systems:
– transcriptions of pronunciation for the pronunciation data category,
– spell check algorithms with pronunciation constraints,
– "phonetic search" wordlists with functions defining phonetic similarity via algorithms such as *Levenshtein distance* or *soundex*;

Machine lexica for spoken language systems (currently always scenario constrained and corpus based):
– orthography-pronunciation mapping for text-to-speech lexica in speech synthesis, in which *orthographic noise* due to homophony and heterography is minimised,
– lexical search and mapping to orthography from output of decoder component in speech recognition in conjunction with a language model,
– translation lexica for speech-to-speech translation,
– resource lexica for generating optimised lexica,
– stochastic language models (essentially wordlists with statistical constraints on contexts of cooccurrence, with *n*-gram, regular grammar or context-free grammar structures.

This heterogeneity makes it somewhat difficult to integrate the requirements for spoken language lexical resources into the generic ISLE framework without considerable backtracking into the basics of computational spoken language lexicography.

H.1.6 Overview

The following sections deal with spoken languge and multimodal lexica, types of lexical information, transcription in corpora and lexica, corpus and lexicon annotation, and formal prerequisites for spoken language lexicon implementation, followed by conclusion and prospects.

H.2 Spoken language and multimodal lexica

H.2.1 Initial clarification of concepts

To provide a systematic framework for the report, an informal and practical implementation-level definition of a monolingual lexical entry is given, excluding frequency (mutual information), etymological and housekeeping meta-information. The definition generalises and simplifies the MILE definition for present purposes. Understanding of the defining concepts is assumed, and will be extended below.

Core lexicon: flat *a 4-column table of the non-compositional signs of a language, with their category, constituent, semantic, and surface properties, with no macrostructural optimizations.* A Core Lexicon may be formalised in terms of attribute-value matrices, and for the purposes of lexicographic software engineering it may be represented and expanded in terms of an Entity-Relationship.

Model (ERM), as in the results of the ISLE Computational Lexicography Working Group (ISLE CLWG). Each of the four fields must, of course, be expanded into groups of sub-fields, corresponding to the Basic Notions and the main data types developed in the ISLE project.

A second level of definition is also needed. This level is envisaged for future ISLE-related activities, is touched on in the ISLE project report, and concerns shared lexical information which is common to sets of lexical entries:

Generalised lexicon: *a hierarchy of definitions of the meanings of categories used in a Core Lexicon.*

Formally, a Generalised Lexicon may be represented by a type subsumption lattice, as in the Unification Grammar paradigm, or from the point of view of lexicographic software engineering as hierarchical Object Oriented Lexicon, as in the DATR paradigm and related approaches.

The ISLE report envisages development of an OOL model at a later stage. The lexica which currently come closest to realising this second level of definition are WordNets and object-oriented or inheritance based lexica. However, the *front matter* of conventional lexica, in which sketch grammars are provided as explanations of the categories used in lexical entries, also has this property in a very practical and simplified form.

H.2.2 General areas of difference between speech and text based lexica

Spoken languages differ from written language lexica at both the core and generalised levels of definition outlined above. This report, like the main ISLE report, concentrates on the first level, and cover differences in structure, content and use, following the broad view of spoken language lexica, including software system lexica, which has already been outlined:

- In structure, spoken language lexica, particularly those constructed for use in spoken language systems, differ from written language lexica in several ways. The most important way is perhaps the need to link lexical entries, via timestamps, to occurrences in corpora, whether for the training of statistical decoder models or for the construction of audio concordances.

- In content, spoken language lexica require information of varying detail about pronunciation, and about the use of lexical items in dialogue contexts and differing pragmatic situations, as well as statistical information. At a future stage, the introduction of *Pragmatic Unit*s, *Prag*U, will be required, extending the <Mu,SynU,SemU> object vector adopted by the ISLE project.
- In use, spoken language lexica have a different deployment spectrum from written language lexica, particularly in the spoken language technologies, as already outlined.

H.2.3 Lesser known domains of Spoken Language Lexicography

Relatively few of the world's approximately 7000 languages are written languages, and construction of lexica for purely spoken, i.e. unwritten languages is perhaps the major single task in descriptive linguistics. A large part of the task is taken up with representing segmental (phonemic) and suprasegmental (prosodic) pronunciation information in the lexicon, and with mapping this to more detailed phonetic representations of pronunciation in actual utterances. In addition, spoken dialogue contexts require the differentiation of different semantic and pragmatic vocabulary fields for representation in the lexicon. Unwritten languages are not intrinsically less complex than written languages: they are supported by complex oral traditions with orally transmitted legal and religious systems, and sophisticated orature (oral "literature").

It is now widely recognised that spoken language is multimodal and not restricted to the acoustic-auditory modality, implying that spoken language systems have to consider "body language" components, including the

- gestural (movements of head, face and limbs),
- postural (configuration of body), and
- proximal (interlocutor distance)

components of communication, as well as the more well-known locutionary components (though the latter are presumably the most complex by orders of magnitude). Until recently, these components have been investigated separately in different disciplines, from choreographythrough anthropological linguistics to the study of the complex sign languages used by acoustically handicapped.

Indirect confirmation of this generalisation of the definition of speech from the acoustic-auditory modality to multimodal communication is provided by the numerous contributions on multimodal spech in the recent events in the LREC and EUROSPEECH conference series; see also [10].
The lexical information required for multimodal speech will therefore be required in the model developed here. However, it is too soon to consider standardization, though initial formalisations, which may form the basis of recommendations at a later date, are available [9].

H.2.4 A note on spoken language genres

Spoken language lexica for system use are almost invariably scenario constrained corpus lexica, while spoken language lexica for direct human use are invariably general language lexica. Scenario constraints correspond largely to the criteria used in the traditional characterization of registers, genres and sublanguages. The range of these spoken language registers, genres and sublanguages is wide, and beyond the scope of this study; it will be sufficient for present purposes to refer to previous traditional studies of genre, register and sublanguages, cf. [7].

In the contexts of the speech sciences, psychology, and spoken language engineering, the most commonly used genre of spoken language is *read speec*h. In anthropological and descriptive linguistics, as well as in ethnolomethodology, conversation analysis and discourse analysis, spontaneous types of dialogue are focussed on. Increasingly, this is also the case for the spoken language technologies, under the influence of development requirements of producing interfaces for natural human-machine interaction.

H.2.5 Basic terminology for spoken language lexica

The basic terminology used in this contribution follows and systematises the usage in previous work in this field, and as far as possible is kept compatible with the work of the ISLE Computational Lexicon Working Group.

Annotation: the enhancement of

- a transcription with symbols paired with timestamps pointing to boundaries or segments in corpus data recordings (labelling); formally, a pair <*label,timestamp*>, where *timestamp* can be a *point* or an *interval*, the *interval* generally being represented by a pair of *point* timestamps <point_i, point_{i+1}>
- written language corpus data by a function mapping descriptive categories into boundaries or segments in the corpus data (tagging, tree-banking).

Core Lexicon: a model of a lexicon as a table with atomic cell entries.

Corpus: a quadruple <*metadata, signaldata, annotations, corpuslexicon*>

Corpus data: collection of (generally digital) audio/video/sensor signal recordings and/or transcriptions of spoken language utterances or hardcopy/scans and/or discrete electronic versions of written language inscriptions.

Corpus lexicon: a set of lexical items (words, idioms) induced from the corpus by the following functions:

- sorting,
- removing duplicates,
- (optionally) stemming, i.e. removing affixes,
- (optionally) lemmatising, i.e. extracting stems as lemmata.

and mapped into a set of types of lexical information.

Generalised Lexicon: a lexicon enhanced by a hierarchy of definitions of categories in the Core Lexicon, constituting a type subsumption or default inheritance hierarchy (cf.Mesostructure).

Lexical entry: a row in the Core Lexicon (or some expanded version of this) representing a vector of types of lexical information (corresponding to the traditional ‘lexicon article’'); see the MILE discussion in the ISLE report.

Lexical data category: a column in the Core Lexicon (or some expanded version of this) representing a type of lexical information contained in lexical entries and corresponding to an Attribute in an Attribute-Value-Matrix (AVM) approach, or, for example, the elements of a Syntactic (or other) Frame in the ISLE model.

Macrostructure: the organization of the entire lexicon (as a table, hierarchy, or hybrid structure); an ordering relation over the rows in a lexical relation table as an search-oriented data structure optimising operation, such as an alphabetic sorting by the orthography, a hyponymy-based tree relation induced over senses of lexical entries, as in a thesaurus.

Mesostructure: the structure of the generalised lexicon, at a level between microstructure and macrostructure; a classification of lexical entries, and an ordering over these classes as a generalisation operation, induced from data category values of lexical entries, and represented as grammatical information in the front matter of a lexicon, or as an inheritance hierarchy in a formal lexicon.

Microstructure: a vector of data categories comprising the types of lexical representation represented in lexical entries.

Modality: a pair of human output (motor gesture) and input (sensory) channels such as <*acoustic*, *auditory*> (e.g. speech),<*gestural*, *visual*>(e.g. gesturing, signing), <*gestural*, *auditory*> (with gestures transduced into sound, e.g. hand-clapping), <*gestural*, *tactil*> (e.g. shoulder-slapping, kissing). In phonetics, speech is also commonly regarded as a specific kind of <*gesture*, *auditory*> modality in which the gestures are restricted to the vocal tract and transduced into sound. Analogously, orthography is a <*gesture*,*visual*> modality in which gestures (typed or handwritten) are transduced into stored traces (inscriptions).

Onomasiological lexicon: a lexicon with a macrostructure optimised for search as a function from sense representations to surface forms; see semasiological lexicon (though there are many other orderings, particularly for multilingual lexica, not covered by these two terms).

Spoken Language Reference Microstructure (ISLE-SLRM) model: a recursively structured vector of data categories (ignoring housekeeping and entry metadata):

Minmal standard model:
<*STRUCTURE*, *INTERPRETATION*>

Standard models:
The standard model may be represented minimally as a pair (shown above), standardly as a quadruple

<*CATEGORY, PARTS, MEANING, SURFACE*>

derived by decomposing STRUCTURE into a pair<*CATEGORY, PARTS*> and <INTERPRETATION> into a pair <MEANING, SURFACE>. SURFACE is, in turn, a pair <$\text{MODALITY}_{\text{acoustic}}$, $\text{MODALITY}_{\text{visual}}$>, and decompositions of finer granularity may be introduced as needed.

Semasiological lexicon: a lexicon with a macrostructure optimised for search as a function from stems to sense representations; see onomasiological lexicon (though there are many other orderings, particularly for multilingual lexica, not covered by these terms).

Submodality: an autonomously organised stream of intonation which modulates a modality (required for representing parallel streams of information such as prosody).

Transcription: a symbolic representation of corpus data; a component of an speech signal annotation.

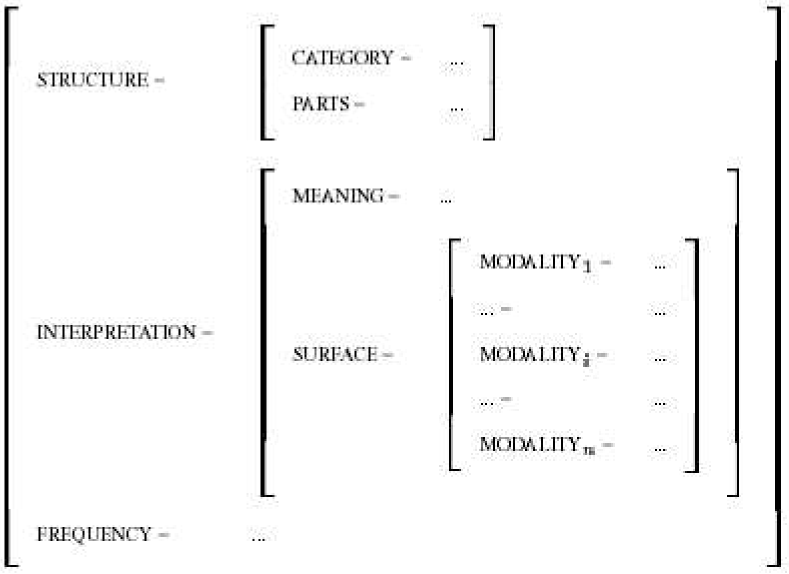
Each element of the microstructure represented by the ISLE-SLRM quadruple needs further decomposition, depending on requirements in specific applications. Additionally, a category for corpus frequency information of different types (isolated frequency, frequency in various contexts such as digrams) is also needed.

The grouping of data categories is derived from the ILEX model [8] and are closely related to the data category specifications of the ISLE Computational Lexicon Working Group, but extended for application to spoken and multimodal lexica. The CATEGORY attribute, in a spoken language lexicon, is very often a statistical function relating co-occurring neighbours in a corpus, but it may also be a function from the lexicon into the corpus which effectively defines a (pre-compiled or on-the-fly) concordance. If the corpus is purely textual the concordance is conventional. However, if the pointers are timestamps relating to an audio or video signal, the concordance is a multi-media concordance (audio and/or video concordance) for human use or statistical system training, For ease of comparison a full version of the ISLE-SLRM model, in somewhat expanded form, is given as a feature structure in Figure 1.

This schema stands for a family of reduced or expanded microstructures in practical instantiations of the model, which depend on actual requirements in specific applications, and represent special cases of the ISLE-SLRM reference model.

For example, the model proposed by Bell & Bird [2] for lexicon metadata definition is triple which may be represented as <*STRUCTURE*, *MEANING*, *SURFACE*> corresponding to a possible instantiation of the standard quadruple described above.

A descriptive linguistic glossary would be adequately modelled by the<*MEANING*, *SURFACE*>, with *MEANING* modelled by a gloss in the description language, and *SURFACE* modelled by a phonemic transcription in the source language.



Grouped representation of microstructure vector in the ISLE-SLRM model as an attribute-value structure. The modalities are ortography, phonetic, gestural etc. (see text).
The PARTS correspond to the Daughters attribute of HPSG-like grammars.

A lexicon in a theoretical linguistic framework, on the other hand, such as the HPSG paradigm, requires a much fuller spelling out of the *STRUCTURE* and *MEANING* attributes.

H.3 Types of lexical information

H.3.1 Generic lexical information

A generic model of syntactic and semantic types of lexical information for spoken language will correspond to the *Basic Notions* of the ISLE CLWG model for written language, presented in the body of ISLE report, and therefore need not be spelled out in this section.

H.3.2 Spoken language specific lexical information

The following are the most important types of lexical information which differ either gradually or categorically from the types of lexical information included in written language lexica:

1. Pronunciation representation in human readable lexica (implementation problems discussed below):

(a) non-standard, highly language-specific adapted orthographies,
(b) International Phonetic Alphabet, currently represented in a wide variety (several dozens) of institute-specific or more-or-less generic fonts,
(c) Alphabets similar to the IPA or encoding the IPA in typewriter-friendly ASCII codes, e.g. the SAMPA alphabet [10].

2. Pronunciation representation in machine readable lexica (this systematisation goes beyond conventional descriptions in the disciplines concerned and generally appears unconventional to practitioners of these disciplines):

(a) orthographic representation supplemented by a grapheme-phoneme conversion component (it is important to note that phonology-orthography relations are highly complex [11], with homophony and homography interaction and pronunciation variants requiring relational specifications of a similar degree of complexity to relational semantic mappings),
(b) statistical characterization of components of pronunciation such as phonemes, diphones, disyllables and larger units by means of stochastic models such as Hidden Markov Models (HMMs), Neural Networks (NNs), Bayesian Networks;
(c) statistical characterization of distributional properties of functional items such as morphemes and words by means of stochastic models of cooccurrence in corpora (digram, *n*-gram models, regular grammars (equivalently: finite state automata), context-free grammars).

3. Gesture (as captured in current gesture transcription systems):

(a) HamNoSys: a gesture representation for sign languages,
(b) FORM: a gesture representation for automatic gesture recognition,
(c) CoGest: a linguistically motivated representation of gesture forms and functions [9].

4. Spoken language specific units corresponding to and generalising the notion of Part of Speech (POS), particularly

(a) interjections, including discourse particles and lexicalised single-word and multi-word hesitation phenomena,
(b) clitic unit formation,
(c) frozen functional units such as function-word sequences,
(d) Stochastic Regular Language Models (e.g. HMMs) and Stochastic Context Free
Language Models [11].

H.4 Transcription in corpora and lexica

The issues of corpus and lexicon transcription, including prosodic transcription, were dealt with in detail in the SAM and EAGLES I and II projects [11], [10] and need only be summarised here.

H.4.1 Transcription in corpus representation

Transcriptions occur in their own right as representations of recorded or informally heard spoken utterances, and with timestamped symbols in annotations (discussed separately below). The kind of transcription used in corpus representations is highly variable and dependent on scenario constraints. The following tentative scale of transcriptions is proposed for corpus representations:

1-tier transcription: Minimally, a phonemic transcription or an orthographic transcription which is regularly related to pronunciation. The latter is perhaps the most widely used form of transcription, both in corpora for spoken language systems, as well as in written language corpora if alphabetic or syllabic orthographies are used. Logographic orthographies (including, for example, arabic and roman numerals) are unsuitable unless a function which maps them into a some pronunciation notation is provided.
2-tier transcription: Preferably, for minority language corpora and especially corpora of endangered language data aligned parallel transcriptions of forms (minimal transcription) and of functional categories (e.g. glosses in a standard description language such as English, French, Russian, Spanish, ...). This kind of transcription is generally referred to as an *interlinear gloss* in descriptive linguistics.
***n*-tier (multi-tier) transcription**: In addition to a 2-tier transcription further aligned tiers with other lexically relevant information, such as:

- non-pronunciation-friendly orthographies,
- prosodic categories,
- morphosyntactic categories,
- semantic categories,
- pragmatic categories.

Any or all of the data categories in the ISLE CLWG microstructure definition can be expected in an annotation tier, in addition to corpus-specific tiers.

H.4.2 Transcription in lexical representation

In addition to any orthographic representation, a transcription system for the instantiation of a Data Category in the microstructure of a lexicon is by definition a *lexical* or *phonological underlying* transcription, not a *phonetic* transcription, which has the function of capturing phonetic detail of the pronunciation of sounds in context. For lexical transcription, the following levels are appropriate, depending on the typology of the language (no distinction will be made here between *prosodic* and *suprasegmenta*l, as distinctions made in the literature are generally idiosyncratic to a particular methodology):

Segmental: Depending on the degree of abstraction in the phonetic and phonological analysis involved in word identification, the following levels may be used:

1. morphophonological or phonotypic: phoneme variation in uniquely identifiable morphological contexts are not marked but generated by rule; e.g. German orthography is morphophonemic Mond /mo:nt/ ‘moon’ - Monde /mo:nd / ‘moons’. For a monolingual lexicon meant for native speakers, who have internalised the syllabification and final devoicing rules, or computational models which have explicit syllabification and final devoicing rules, no distinction is necessary,
2. phonemic: phonetic variation between the alloophones of phonemes within a word (such as the approximant and vocalised allophones of /r/ or the aspiration of voiceless plosives in English), without consideration of word-internal morphological information,
3. phonetic: variations in phonemic or phonetic detail which may be required for capturing phonostylistic variation of pronunciation in different styles and registers.

The mesostructure of a spoken language lexicon (the front matter of a book lexicon, the rules or inheritance hierarchies of a formal lexicon) contains generalizations to supplement the explicit information in lexical entries:

- *morphophonemic rules* for specifying phoneme variation in morphophonemic contexts,
- *allophonic rules* for specifying allophone variation in phonemic contexts,
- *phonetic detail* rules for specifying dialectal and sociolectal variation,
- *supralexical rules* (traditionally known as “postlexical rules” in Generative Phonology and its descendants) for specifying pronunciation variation in larger syntactic and phonostylistic contexts.

Suprasegmental (prosodic): In languages which have lexical prosody such as word stress or tone, with various distinctive and morphological functions, provision must be made for representing this. The main kinds of lexical prosody are the following:

*stres*s, as in Dutch, English, German;
*pitch accen*t, as in Japanese;
*tonal accen*t, as in Swedish;

*register ton*e, as in very many African languages;
*contour ton*e, as in very many South East Asian languages.

The main lexical and morphological functions of lexical prosody are, illustrated with German examples (for clarity: stress indicated by upper case, standard initial upper case characters not used):

*distinctiv*e: *TenOR* ‘tenor’ (singer) - *TEnor* ‘tenor’ (gist)
*inflectiona*l: *DOKtor* ‘doctor’ - *dokTORen* ‘doctors’
*derivationa*l: *TElefon* ‘telephone’ (noun) - *telefoNIEren* ‘telephone’ (verb),
*compoundin*g*: ÜBersetzen* ‘cross over’ (verb), *¨berSETZen* ‘translate’ (verb).

These functions occur with all types of lexical prosodic forms. Further relatively standardized information about prosodic transcription, with examples from 20 languages, can be found in [12].

It is clear that a structured notion of *Surface Uni*t, *Surf*U, will be required in future work in order to accommodate lexical information about pronunciation, including lexical prosody, and their complex interrelations with gestural and related lexicalised multimodal information.

H.4.3 Implementation of transcriptions

H.4.3.1 CRITERIAL LEVELS

The following sections pertain mainly to the implementation of lexica for human use, and not primarily to the implementation of system lexica, which are very product-specific. In implementing transcription systems, at least the following four-way level distinction is needed:

1. *transcription category* (e.g. a voiceless alveolar plosive, a high tone),
2. *transcription id* (e.g. the code numbers in the Esling coding of the IPA, or the SAMPA ASCII encoding of the International Phonetic Alphabet, cf. [10], or in the Unicode conventions),
3. *transcription symbol* (e.g. /t/ for a voiceless alveolar plosive phoneme, ´ for a high tone),
4. *transcription font* (e.g. the actual glyph design for particular symbols) with properties such as *serif* or *non-seri*f.

H.4.3.2 TRANSCRIPTION CATEGORY

The level of transcription category is rather stable for segmental transcriptions, and is detailed in the handbook of the International Phonetic Association [13] and in practically any leading phonetics textbook such as [6].
For prosodic transcriptions this is not the case. The IPA provides a set of categories for prosodic transcription, but other sets are in use. The variation is too great to be detailed here, but several examples can be found in [12].

H.4.3.3 TRANSCRIPTION ID

The level of transcription id has not completely stabilised, as there are a number of codings in current use, as already mentioned. It may be supposed that the Unicode conventions will be adopted as soon as adequate rendering engines for Unicode become available, which is currently not the case. For this reason, other encodings are commonly used.

H.4.3.4 TRANSCRIPTION SYMBOL

At the level of symbol choice, the situation is also rather stable, since most symbols in current use have been fixed by the International Phonetic Association since the late 19th century.

There is still some variation in different traditions, however, for example between /y/ in one US phonetic tradition, and /j/ in European phonetic traditions, for a palatal approximant, Although this distinction is rather trivial, nevertheless it can lead to misunderstandings. Consequently the recommendations of the International Phonetic Association are preferred.

In spoken language technologies and in general lexicographic practice conventions such as accent diacritics over words for stress or tones, or capitalisation of stressed syllables, are found.

H.4.3.5 TRANSCRIPTION FONTS

There are many fonts available for encoding segmental symbols, from common orthographic fonts to specific implementations of the IPA symbols. Font implementations are very platform-specific, and easy font conversions are not possible.

Simplifying away from a number of issues which are not immediately relevant for present purposes, there are two basic font technologies which affect the usability and interoperability of fonts, word processor outline font technology such as TTF, and the Metafont glyph function technology used by TEX systems, mainly under UNIX.

The most well-known word processor IPA font implementations are:

- IPAkiel, based on the IPA chart as defined at the Kiel Convention in 1993.
- SIL (Summer School of Linguistics) fonts, of which there are many, some being local variants, and perhaps the most well known being the Doulos series.
- Lucida, which maps into Unicode, and will possibly be the preferred font for this purpose when full Unicode rendering engines for phonetic symbols and complex diacritic stacking are available; current engines do not have this capability, consequently enforcing non-standard solutions.

For the WYSIWYG oriented fonts, even with similar implementations, the functions which map ids to symbols and their glyphs, and which map keyboard combinations to ids and thence to symbols and their glyphs) vary greatly from one font implementation to another.

The most well-known implementations of IPA fonts for LATEX are:

- WSUIPA from Washington State University,
- TIPA from University of Tokyo.

The WSUIPA font was used in the production of the EAGLES Spoken Language Working Group handbooks [11] and [10].

H.5 Corpus and lexicon annotation

H.5.1 The definition re-visited

As defined in the glossary, an annotation in the context of spoken language context is:

Annotation: the enhancement of a transcription with symbols paired with timestamps pointing to boundaries or segments in corpus data recordings (labelling); formally, a pair <*label, timestamp*>

where *timestamp* can be a *point* or an *interval*, the *interval* generally being represented by a pair of *point* timestamps <$point_i$, $point_{i+1}$>

The theoretical foundations of this concept of annotation are due to event logic based phonologies, as in the Event Phonology of Bird & Klein, cf. [3], [5]. The Time Map theory of Carson-Berndsen [5] introduced extensions to phonetic levels and applications to spoken language processing, with finite state models for the event logic theory.

Annotation of spoken language data has been extensively dealt with, also in previous European project work, cf. [11] and [10]. Currently the most influential approach is by Bird & Liberman [4], who generalised the notion of annotation by means of the construct *Annotation Graph* in order to encompass known kinds of signal and text annotation.

Strictly speaking, the inclusion relation

$$Transcription \subset Annotation$$

holds, since maximally the beginning and end of any transcription are informally synchronised with the beginning and end of the speech signal. This is not a very useful idea, however, as the synchronisation is too fuzzy in general to be machine processable, and in any given case a signal recording may not actually exist.

H.5.2 Acquisition of spoken language lexical information

Work in the spoken language technologies, and modern hyperlexicon applications such as audio and video concordances, presuppose the availability of carefully annotated spoken languagedata. Many tools for automatically, semi-automatically and manually annotating audio signal data are available. Of primary importance are the manual tools, since the final quality criterion for accuracy (not necessarily consistency!) of annotations is the human annotator.
There are many proprietary and locally developed and used tools for audio annotation, and not many at all for video annotation. These developments are not available for standardization, and have not developed into state of the art freeware or open source tools, and consequently, they will not be considered here.
Currently, freely available and rather widely used tools for this purpose are the following:

1. Praat, a comprehensive freeware toolset for annotation and experimentation in phonetics and speech technology, developed since the early 1990s at the University of Amsterdam phonetics laboratory by Paul Boersma and David Weenink.
2. Transcriber, a tool originally developed for radio broadcast annotation, developed at ICP Grenoble by Claude Barras and Edouard Geoffrois, and ported to other environments at LDC, U Pennsylvania.
3. esps/waves+ ("Xwaves"), a proprietary library and GUI developed by Entropic in Cambridge, UK, and not maintained or generally available since its purchase by Microsoft Corp. a number of years ago (though a change in this policy is apparently under discussion).
4. WaveSurfer, a library and GUI for speech annotation and analysis developed at KTH, Stockholm.
5. TASX, an open source workbench for video and audio annotation developed by Jan-Torsten Milde at Universit¨ at Bielefeld and implemented in Java.

Currently the most widely used audio tools, in research, development, annotated resource creation and teaching environments, are Praat and Transcriber.
In comparison with lexicon acquisition work on written language corpora, spoken language lexicon acquisition is relatively impoverished. Although the automatic construction of stochastic models for speech recognition is very highly developed, this is not true of the analysis of collocations, vocabulary fields and related activities characteristic of corpus linguistics and of natural language processing in general.
In addition to these activities, the following areas of generalisation over corpora are on the horizon, represented by a scattering of studies, and it may be expected that these will increase in importance as spoken language input and output devices become more widely used and adaptability and portability requirements increase in importance:

1. resource adaptation in spoken language technology, for example for permitting a generic speech recognition or speech synthesis application to be used by a wider range of users,
2. annotation graph collation by fuzzy operations over near-simultaneous points and overlapping intervals in time,
3. syntagmatic hierarchy induction in order to create phonotactic, morphotactic and phrasal grammars from data,
4. paradigmatic induction of class hierarchies over lexical items for use in compact and robust inheritance hierarchies.

Effectively, these are Machine Learning (ML) applications which are gradually being transferred from other areas of language processing to spoken language, and will not only make the reusability and inter-operability of spoken language resources more feasible but will also enable resources to be related to and benefit from theoretical linguistics, and vice versa.
Finally, it may be noted that "lexicon acquisition" in general relates to the instantiation of predefined microstructures from corpus data. The notion may be generalised, however, to the process of defining lexicon microstructures by means of generalisation and disjunctive abbreviation procedures (e.g. for alternatives in lexical fields).

H.6 Formal prerequisites for spoken language lexicon implementation

H.6.1 Towards an integrated generic lexical model for spoken language

The ISLE approach to the characterization of lexicographic representations in terms of *Basic Notions* and its application to lexicographic software design in terms of an Entity-Relationship-Lexicon Model (and later an Object-Oriented Lexicon Model) is easily adapted to the needs of Spoken Language Lexicography, as noted in previous discussion in this report. However, in order to fulfil the needs of Spoken Language Lexicography in detail, a precise distinction between stages and levels of formalisation is required (with appropriate declarative and procedural specifications at each level which will not be discussed further here):

Lexicon characterisation: The linguistic specification of requirements for lexicon structure, types of lexical information, and generalisation over lexical subregularities.
Lexicon formalisation: The mathematical specification of structures such as attribute-value-matrices, feature vectors, entity sets, etc., required for defining the structure of lexica.
Lexicon engineering design: The specification of abstract data structures for software engineering purposes, such as Entity-Relationship or Object-Hierarchy models.

This distinction is somewhat maximalistic, but in principle each of these levels need to be addressed to a greater or lesser degree during the lexicon design process, and before a lexicon is implemented. The most important level is *lexicon characterisatio*n. Traditional lexicogaphy is very much a practical art, and many features of this art have been transferred to computational lexicography. The preceding sections of this report constitute a preliminary specification of issues for linguistically based lexicon characterization within the domain of Spoken Language Lexicography. The concepts of microstructure (especially the ISLE-SLRM model of lexicon microstructure), macrostructure and mesostructure were introduced earlier as a basic generic starting point for defining a wide range of varieties of spoken language lexicon more precisely, and for integrating the notion of spoken language lexicon into the generic approach adopted by the ISLE Computational Lexicon Working Group.
Summarising the discussion of previous sections, the structure of the basic reference lexicon is
thus a triple:

<*microstructure, macrostructure, mesostructure*>

The microstructure, which corresponds to the notion of *lexicon model* or *lexical entry model* in much text-based computational lexicographic work, defines the structure of the vector which represents specific lexical entries *LI* means “lexical information” and *TLI* means “type of lexical information”, usually expressed as attribute-value pairs):

$$<LI_{TLI1}, \ldots, LI_{TLIn}>$$

In this ISLE-SLRM reference model, microstructures are intended to be fully inflated, with no macrostructurally motivated disjunctions or substructures of the kind to be found in conventional alphabetic semasiological dictionaries. The reason for this is to clarify the fact that disjunctions and tree-structured lexical entries implicitly express generalizations over more primitive structures.

The generic ISLE-SLRM quadruple model (decomposed further to the required degree of granularity) is mapped on to specific lexica by means of a set of generalisation operations which define mesostructures and the macrostructures of the specific lexica for the purpose of different strategies of lexical access, and the development of particular *views* or *indexings* on lexical databases which are associated with different access strategies, and which are particularly important for the development of multimedia hyperlexica for multilingual and multimodal information.

The main operations for mappping a non-generalised lexicon into a specific lexicon view are:

1. A grouping operation for local alternatives in pronunciation in spelling data categories which are unrelated to other data categories (analogous to the grouping of readings or senses). A simple example is the /aɪð□- ið□/ pronunciation variation of English *eithe*r.
2. A distributed disjunction operation (due to Krieger & Nerbonne), for example for relating linked local alternatives in pronunciation which are linked to alternatives in spelling (or other data categories), or in specifying morphological syncretisms. This operation is usually represented in lexicography by postulating separate lemmata when the disjunctions include syntactic or semantic categories. However, since each primitive lexical entry in the ISLE-SLRM quadruple (with an appropriate level of further decomposition) is inherently separate from all others, lemma groupings are by default left over after the groupings of alternatives and distributed disjunction operations have been applied.
3. Abstraction of distributed disjunctions into a type or default class hierarchy.

4. Definition of macrostructure ordering relations over lexical entries based on relations between the fields of lexical entries. Macrostructures which differ from the basic table structure are generally application specific, and defined for optimization in lexical search, whether by human or machine. Macrostructures, therefore, contrast with other generalizations, which may be said to be *declarative*e, in that they - while they are, seen in isolation, declarative - have a *procedural* motivation. Examples of procedurally motivated macrostructure orderings are:

    - Language ordering: tabular ordering with access via one language or another in a bilingual or multilingual lexicon.
    - Onomasiological ordering: tree graph induced over sense terms and representing a hyponym-hyperonym taxonomy or a meronymy (as in a thesaurus).
    - Semasiological ordering: alphabetical ordering over the orthography field (classically, 'the dictionary') or - in spoken language lexica - by the pronunciation field(s).
    - An unnamed ordering: alphabetical ordering over reversed pronunciation fields (rhyming lexica).
    - Rank ordering for defining idiom dictionaries as against word dictionaries or morpheme dictionaries, etc.
    - Selection by sublanguage field for technical and terms and other collocationally restricted items for specialised sublangage dictionaries.
    - Speech recognition ordering: over representations of pronunciation.
    - Speech synthesis ordering (text-to-speech): as semasiological ordering, except in general with highly reduced microstructures.
    - Speech synthesis ordering (concept-to-speech): as onomasiological ordering, but in principle with no intervening orthographic representation.

In traditional lexicography, macrostructure orderings such as these define different specific book lexica. In the context of lexical databases, these macrostructure ordering relations are implemented as database views. In hyperlexica, the macrostructures are defined as alternative super-imposed hyperlink structures.

Any resource archive format will need to be at least "virtually" reconstitutable into the ISLE-SLRM quadruple format with a decomposition of appropriate granularity in order to be able to map the format on to the different microstructures required in speech technology and human readable dictionary publication.

H.7 Conclusions and prospects

H.7.1 Integration of multilingual multimodal lexica

It has been shown that by returning to some basic concepts the apparent heterogeneity in spoken language lexicography can be related to the basic notions of the ISLE approach; the realisation of spoken language lexica themselves uses very different lexicographic techniques, however.

The ISLE-SLRM (Spoken Language Reference Microstructure) quadruple model, at whatever degree of decompositional granularity, like generalised models of lexical entries used in some theoretical linguistic frameworks, has the property that notions of headword and lemma are not basic, but the result of the application of operations of *mesostructural generalisation* and *macrostructure optimization*. In a model of this kind, multimodal lexical representations may be integrated using similar mechanisms to those used for text-oriented lexica.

The results of procedure followed in this contribution are thus similar in spirit to the results outlined by Bell & Bird [2], except that

- they based their study on an examination of 55 printed lexica, whereas this contribution is additionally based on varied practical experience in manufacturing complex spoken language lexica,
- heir abstract data model is more restricted, and turns out to be a special case of the ISLE-SLRM quadruple model,
- they do not systematically consider macrostructural optimizations or mesostructural generalizations of the lexicon (though one of their main points is that lexica are extremely inhomogeneious).

There has been no explicit discussion of metadata in relation to the ISLE-SLRM model because of the need to clarify requirements and design issues before moving on to specific metadata proposals for implementing archives and dissemination portals.

However, the ISLE-SLRM model has been designed in such a way as to assist metadata design, and it is proposed that the ISLE-SLRM model, with the associated concepts of macrostructural optimization operations and mesostructural generalisation operations is a suitable foundation for the definition of lexical metadata for spoken language lexica.

H.7.2 Realisation with XML technologies

The formal character of an attribute-value structure is common to linguistic feature structures, to the ISLE-SLRM model, and to XML tree structures, facilitating portability of structures from one methodology to another.

In the context of spoken language lexicography, a number of theoretical problems remain with XML, though in one way or another the structures discussed here may be represented in XML (and RDF). The following points are relevant for the generation of the varieties of macrostructure required for different kinds of spoken language lexicon:

1. XML has no well-defined formal semantics beyond the assignment of tree-graphs; pointer structures, for example, are semantic structures and not definable within the context-free syntax.
2. For other abstract data structures, *ad hoc* definitions are required.
3. A specific example of an *ad hoc* solution to a well-known problem is the case of tables (familiar from the LATEX and HTML table models): if the XML tree is row-based, there is no well-defined concept of column; the column is stipulated in an informal semantics for the tree. The same holds vice versa. If it is accepted that XML is specifiable by context-free rules, the proof of this is obvious: a table has the structure $a^n\ b^n\ c^n\ \ldots$, which is clearly not context-free but context-sensitive (in fact, an indexed language). An actual implementation for the purpose of system use in spoken language technologies (or browser construction) must take this into consideration.
4. The issue becomes correspondingly more complex with recursive tables.
5. Another specific example of *ad hoc* solutions, from the point of view of the syntax of XML, is the enhancement of XML with pointers. The internal syntax of pointers is tree-structured, but their semantics is that of variables with arbitrary values over positions in documents, which may be used to construct structures of arbitrary complexity.
6. it is unlikely that the transformations required by spoken language system technologies will be amenable to the procedural components of the XML technologies such as XSLT.

These points pertain to the projection of the ISLE-SLRM and will no doubt be resolved when an adequate semantics for XML is available. Of course, in practice many of these issues can be relativised: for instance, with small tables of predefined dimensions the complexity problem for tables reduces to linear complexity. Currently it is sufficient to note them for future discussion and to illustrate the necessity for pragmatic strategies of operationalising computational lexica for spoken language.

H.7.3 Final remarks

The ISLE approach is based on the following key ideas which are of central relevance to the present report:

- bottom-up,
- based on well-defined *Basic Notion*s,
- addresses formalisation issues in terms of an Entity-Relationship model.

These features are easily extendable to the incorporation of Spoken Language information, including information about multimodal dimensions of speech, into the family of ISLE recommendations.
A number of systematisations required as prerequisites to the standardization of Spoken Language resources, beyond the recommendations of [11] and [10], are introduced. In particular the introduction of notions *Pragmatic Uni*t, *PragU* and *Surface Uni*t, *Surf*U, the latter subsuming Pronunciation and other communication modalities, was proposed.

A number of issues are not mature enough for recommendations to be provided at this stage:

- application of temporal logic and event logic to the specification of multi-tier lexical prosodic representations (note that these representations are closely related to the lexical temporal and event semantics of verbs of movement and action) cf. [3], [5], and contributions in [14],
- definition of a notion of *Pragmatic Unit* or *Prag*U, and further integration of text-oriented lexicography and Spoken Language Lexicography,
- incorporation of results of ongoing research on multimodal (gesture, posture, proximal) information into lexical inventories,
- extension of existing approaches to lexical metadata to spoken language,
- development of a generic declarative lexicon representation which will permit the compilation of different application-oriented and access-optimised views for both human and machine deployment,
- results on formalisation of lexical structure [1], [14] as a prerequisite to the development of abstract data structures for lexical software engineering,
- resolution of practical representation issues, such as font standardization and XML/RDF specifications of lexical structures.

In these areas, resource systematisation and standardization work is continuing,and the open issues require continuous attention in order to project the consensus oriented goals of the ISLE project into future research and development using shared multilingual multimodal lexical resources.

References

[1] Leila Behrens and Dietmar Zaefferer, editors. *The Lexicon in Focus: competition andConvergence in Current Lexicology*. Lang, Bern, 2002.

[2] John Bell and Steven Bird. A preliminary study of the structure of lexicon entries. LDC, Philadelphia, 2000.

[3] Steven Bird. *Computational Phonology: A Constraint-Based Approac*h. Cambridge University Press, Cambridge, 1995.

[4] Steven Bird and Mark Liberman. A formal framework for linguistic annotation. *Speech Communicatio*n, 33(1-2):23–60, 2001.

[5] Julie Carson-Berndsen. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition.* Kluwer Academic Publishers, Dordrecht, 1998.

[6] John Clark and Colin Yallop. *Introduction To Phonetics And Phonology, 2nd edn.* Black-well, Oxford, 1995.

[7] Penelope Eckert and John R. Rickford, editors. *Style and Sociolinguistic Variation.* Cambridge University Press, Cambridge, 2002.

[8] Dafydd Gibbon. Compositionality in the English inheritance lexicon: English nouns. In Leila Behrens and Dietmar Zaefferer, editors, *The Lexicon in Focu*s, pages 145–185. Lang, Frankfurt a. Main, 2001.

[9] Dafydd Gibbon, Benjamin Hell, Karin Looks, and Thorsten Trippel. Formal syntax of gesture: Cogest-2.0. Modelex technical report (dfg forschergruppe texttechnologie), Universität Bielefeld, February 2003.

[10] Dafydd Gibbon, Inge Mertins, and Roger Moore. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation.* Kluwer Academic Publishers, Dordrecht, 2000.

[11] Dafydd Gibbon, Roger Moore, and Richard Winski. *Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter, Berlin, 1997.

[12] Daniel Hirst and Albert Di Cristo, editors. *Intonation Systems: A Survey of Twenty Language*s. Cambridge University Press, Cambridge, 1998.

[13] I.P.A. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabe*t. Blackwell, Oxford, 1999.

[14] Frank van Eynde and Dafydd Gibbon, editors. *Lexicon Development for Speech and Language Processing.* Kluwer Academic Publishers, Dordrecht, 2000.