Project ref. no.	IST-1999-10647
Project title	ISLE Computational Lexicons Working Group

Deliverable status	Public
Contractual date of delivery	December 2000
Actual date of delivery	February 2001
Deliverable number	D2.1-D3.1
Deliverable title	Survey of Major Approaches Towards Bilingual/Multilingual Lexicons
Type	**Report**
Status & version	Pre-final
Number of pages	238
WP contributing to the deliverable	WP2-WP3
WP / Task responsible	Nicoletta Calzolari, Ralph Grishman, Martha Palmer
Author(s)	Nicoletta Calzolari, Ralph Grishman, Martha Palmer, Sue Atkins, Nuria Bel, Francesca Bertagna, Pierrette Bouillon, Bonnie Dorr, Christiane Fellbaum, Dafydd Gibbon, Nizar Habash, Elke Lange, Sabine Lehmann, Alessandro Lenci, Susan McCormick, Jock McNaught, Antoine Ogonowski, Joseph Pentheroudakis, Steve Richardson, Gregor Thurmair, Lucy Vanderwende, Martha Villegas, Piek Vossen, Antonio Zampolli
EC Project Officer	Brian Macklin
Keywords	Computational lexicons, multilinguality, lexical resources, machine translation, standards
Abstract (for dissemination)	The ISLE Deliverable D2.1-D3.1 from the ISLE Computational Lexicons Working Group is based on previous EAGLES work on monolingual lexicon standards, and presents a survey of the major existing bilingual and multilingual lexical resources. Each resource has been investigated to determine its lexical structure and how it encodes cross-language relationships. The survey documents the results in a format that allows easy comparison of the lexical mechanisms employed by each lexicon considered. The emphasis of the survey is on the semantic level of description as benefits a multilingual perspective, however other levels of lexical information are also taken into account, where these have a bearing on interpretation and on cross-language mapping. This work was undertaken to enable the subsequent identification of a set of basic notions needed to describe the multilingual level, together with other notions that may be recommended for particular purposes or languages.

ISLE Computational Lexicons Working Group

Deliverable D2.1-D3.1

Survey of Major Approaches Towards Bilingual/Multilingual Lexicons

February 2001

Responsible Authors

Nicoletta Calzolari	Istituto di Linguistica Computazionale, CNR, Pisa, Italy Consorzio Pisa Ricerche, Pisa, Italy
Ralph Grishman	New York University, New York City, NY
Martha Palmer	CIS Department, University of Pennsylvania, Philadelphia, PA

Authors

[1]*Sue Atkins*, [2]*Nuria Bel*, [3]*Francesca Bertagna*, [4]*Pierrette Bouillon*, [5]*Bonnie Dorr*, [6]*Christiane Fellbaum*, [7]*Dafydd Gibbon*, [5]*Nizar Habash*, [8]*Elke Lange*, [4]*Sabine Lehmann*, [9,16]*Alessandro Lenci*, [10]*Susan McCormick*, [11]*Jock McNaught*, [12]*Antoine Ogonowski*, [13]*Joseph Pentheroudakis*, [13]*Steve Richardson*, [14]*Gregor Thurmair*, [13]*Lucy Vanderwende*, [2]*Marta Villegas*, [15]*Piek Vossen*, [9,16]*Antonio Zampolli*

1:Word Trade Centre, UK. 2:GilcUB, Barcelona, Spain. 3:Consorzio Pisa Ricerche, Pisa, Italy. 4:ISSCO, University of Geneva, Switzerland. 5:University of Maryland, UMIACS, USA. 6:Psychology Dept., Princeton University. 7:Universität Bielefeld, Germany. 8:SYSTRAN. 9:Università di Pisa, Italy. 10:SAP AG. 11:UMIST, Manchester, UK. 12:LexiQuest, Paris, France. 13:Microsoft, Redmond, USA. 14:Sail Labs, Munich. 15:Sail Labs, Antwerp. 16:Istituto di Linguistica Computazionale, CNR, Italy

Summary

Preface - The EAGLES/ISLE Enterprise

The ISLE project is a continuation of the long standing EAGLES initiative (Calzolari, Mc Naught and Zampolli, 1996), carried out through a number of subsequent projects funded by the European Commission (EC) since 1993. 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 now under the Human Language Technology (HLT) programme as ISLE, since January 2000. ISLE stands for *International Standards for Language Engineering*, and is carried out in collaboration between American and European groups in the framework of the EU-US International Research Co-operation, supported by NSF and EC ISLE was built on joint preparatory EU-US work of the previous 2 years towards setting up a transatlantic standards oriented initiative for HLT. Quite recently we also have some Asian involvement, because of their interest in the initiative and the relevance of lexical standards.

The objective of the project is to support HLT R&D international and national projects, and HLT 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 have actively participated in the definition of this initiative and have lent invaluable support to its execution. Moreover, the initiative is a direct result of a series of recommendations made to the EC over several years. There is a recognition that standardisation 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* standardisation 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 is being conducted with a view to providing the foundation for any future recommendations for International Standards that may be formulated under the aegis of ISO.

The current ISLE project (see http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm) targets the three areas of *multilingual computational lexicons*, *natural interaction and multimodality* (NIMM), and *evaluation of HLT systems*. These areas were chosen not only for their relevance to HLT but also for their long-term significance.

- For *multilingual computational lexicons*, ISLE aims at: 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 standardised evaluation procedures for lexicons.

- For *NIMM*, a rapidly innovating domain urgently requiring early standardisation, ISLE work is targeted to develop guidelines for: the creation of NIMM data resources; interpretative annotation of NIMM data, including spoken dialogue in NIMM contexts; annotation of discourse phenomena, and meta descriptions of multimodal language resources.

- For *evaluation*, ISLE is working on: quality models for machine translation systems; and maintenance of previous guidelines - in an ISO based framework (ISO 9126, ISO 14598).

Three Working Groups, and their sub-groups, carry 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 will be widely disseminated and published, after due validation in collaboration with EU and US HLT R&D projects, National projects, and industry.

The following document presents the results of the first phase of activities of the Computational Lexicon Working Group (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 will be the focus of the others WPs 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.

1 The Computational Lexicon Working Group: an Overview

1.1 Standard 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 will allow 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 have already become *de facto* standards. Standards are not of interest if they are not actually used. 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. This is a very good measure of the impact – and of the need – of such standardisation 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, 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, Leech and Wilson, 1996). 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 the creation of a really large infrastructural platform of harmonised lexicons in Europe, sharing the same model.

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, standardisation 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.

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 current NSF project XMELLT on multi-words for multilingual lexicons will provide valuable input to ISLE.

With no intent of imposing any constraints on investigation and experimentation, the current 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. Therefore, 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 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 proposals.

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 reimplementing what many others have already worked on. Useful indications of *best practice* will therefore come to researchers as well as resource developers. This is the only way our discipline can really move forward.

Finally, one of the targets of standardizaton, 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) in the field of computational lexical semantics and multilingual lexicons, so that synergies will be thus enhanced, commonalities strenghtened, and resources and findings usefully shared. In other terms, the process of standard definition undertaken by the CLWG, and by the ISLE enterprise in general, 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.

1.2 EAGLES methodology

The basic idea behind EAGLES work is 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 being 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 harmonisation must take into account the needs and nature of the different major contemporary approaches and the requirements of different applicative systems and components. EAGLES is also drawing strong inspiration from the results of major projects whose results have contributed to advancing our understanding of harmonisation issues.

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 standardisation 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 harmonisation 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.;
- Testing and validating preliminary proposals;
- Feasibility studies for less mature areas;
- Suggesting actions to be taken for a stepwise procedure leading to the creation of multilingual reusable resources, elaboration of evaluation methodologies and tools, etc.

1.3 The Survey phase

Following the well established EAGLES methodology, the first priority of the CLWG in the first phase of the ISLE project was to do a wide-range survey of bilingual/multilingual (or semantic

monolingual) lexicons, so as to reach a fair level of coverage of existing lexical resources of different types.

This phase is a preliminary and yet crucial step towards the main goal of the current CLWG, i.e. the definition of the "Multilingual ISLE Lexical Entry" (MILE). 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. This is the main focus of the second year of the project, the so called "recommendation phase", where the main objective is proposing consensual Recommendations/Guidelines. Since a substantial part of the basic notions for MILE 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 have 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. To help accomplish this aim, we need to investigate how lexical information is treated in existing monolingual/multilingual dictionaries. The Survey presented in the following chapters of this document covers the survey part of both WP2 and WP3[1] of the ISLE Workplan.[2]

The survey of existing lexicons 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 will form the reference landmarks for the propositive phase of standard design. A number of multilingual applications has been considered as a starting point for both phases, providing 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 are implementable.

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 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 multilingual lexical resources:

- Machine Translation (MT)
- Cross-Language Information Retrieval (CLIR)
- Cross-Language Information Extraction
- Multilingual Language Generation

[1] This merging of WP2 and WP3 was proposed by the project and agreed by the project officer, as stated in the first semestrial report. The final results of WPs 2 and 3 will also constitute one deliverable.

[2] A few American surveys are still expected, due to a late start of the project on the American side. Some Asian surveys are also expected. The current Survey is therefore to be considered still a pre-final version.

- Multilingual Authoring
- Speech-to-Speech Translation
- Multilingual Summarisation

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.

In the preparation of the Survey, i) to facilitate the identification of basic notions and the comparison of surveyed resources, and ii) 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 (see section 2), and

2. to identify a small number of major categories of cross-lingual lexical phenomena that could be used to focus the survey (see section 5). These categories are not intended to be complete, but rather to provide the necessary bootstrap to the propositive 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.

Each summary of a particular bilingual/multilingual or semantic lexicon would in principle include:

1. a description of the surveyed resource (on the basis of the common grid);

2. possibly, for one or two examples from the cross-lingual lexical phenomena, an explanation of how these examples are handled by this lexicon. In the case of semantic lexicons (e.g. SIMPLE or WordNet), the summarizer would separately describe the mapping onto language-independent conceptual levels.

The principle guiding the elicitation and proposal of MILE basic notions in the next phase, based also on the investigation of how lexical information is treated in existing multilingual dictionaries, will be, 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. The work of gathering descriptions and characterisations of multilingual lexical phenomena from a set of major existing lexicons, systems, dictionaries, etc., will provide better ground to then decide what is needed, what can be agreed on, what can be integrated in a unitary MILE, what is lacking or needs formalisation, and so on.

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.

Natural language meaning has always been thought of as one of the hardest problems for standardisation. 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 Summarisation. As this trend consolidates, the need of a protocol which helps normalise 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 will pay particular attention in the second phase, and we are currently examining the extension of the EAGLES guidelines in these and other areas to propose a broad format for multilingual lexical entries which should be of general utility to the community.

In the previous EAGLES work on Lexicon Semantics the following technologies were surveyed to determine which types of semantic information were most relevant:

- Machine Translation
- Information Extraction
- Information Retrieval
- Summarisation
- Natural Language Generation
- Word Clustering
- Multiword Recognition + Extraction
- Word Sense Disambiguation
- Proper Noun Recognition
- Parsing
- Coreference

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[3]:

- BASE CONCEPTS, HYPONYMY, SYNONYMY: all applications and enabling technologies
- SEMANTIC FRAMES: MT, IR, IE, & Gen, Pars, MWR, WSD, Coref

[3] The various abbreviations stand for: MT: Machine Translation, IR: Information Retrieval, IE: Information Extraction, Gen: Generation, Pars: Parsing, MWR: Multiword Recognition, WSD: Word Sense Disambiguation, Coref: Coreference, Word Clust: Word Clustering, PNR: Proper Nouns Regognition, SUM: Summarisation.

- 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. For this reason, as also stated in the Technical Annex, 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 survey. One very interesting possibility seems to be to complement WordNet-style lexicons with the SIMPLE design, thereby trying to get at a more comprehensive and coherent architecture for the development of more comprehensive semantic lexical resources.

MILE will also include previous EAGLES recommendations for other layers. We will evaluate the usefulness of these other layers in the multilingual perspective, e.g. for the MT and CLIR tasks. We will therefore have to analyse whether existing EAGLES recommendations, or existing lexicon models, with respect to the agreed basic notions, comply with the requirements of a multilingual perspective. Differently from previous levels of description, for the multilingual level it will however most probably appear 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 will have also to discover areas of deficiency, and highlight areas in need of further analysis. The same is true of applications: for most/some of the already existing 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.

2 Lexical information in bilingual resources

The preliminary phase of our work has been dedicated to drawing up a list containing the information usually present in various linguistic resources. A first list, proposed by Sue Atkins, essentially concerned the information present in traditional dictionaries, and it has been integrated with more detailed morphosyntactic, syntactic and semantic information, which might be available in existing computational lexicons and machine-readable dictionaries.

The following template has been used as a general grid to evaluate the content and structure of the surveyed lexical resources, verifying if the information is available and extractable and 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 is obviously not intended to be complete, since it is expected that new items might be introduced.

Explanation of abbreviations used in the table below:	**SL** *source language*	**TL** *target language*	**dec** *decoder*	**enc** *encoder*

Table 1: Lexical Information in Bilingual Resources

	Entry component	**Information content**	**Mode**	**Function**
1	Headword	lexical form(s) of the headword: how the headword is spelt	SL	Helps both SL and TL users find the information they are looking for
2	Phonetic transcription	how the headword (or variant form etc.) is pronounced (in *International Phonetic Alphabet*)	IPA	Helps user pronounce the word correctly
3	Variant form	alternative spelling of headword or slight variation in the form of this word	SL	helps both types of user find the information they are looking for
4	Inflected form	other grammatical forms of the lemma (headword)	SL	helps dec user find the information they are looking for helps enc user use the word correctly
5	Cross-reference	indication of another headword whose entry holds relevant information, or some other part of the dictionary where this may be found	code	helps both types of user find the information they are looking for, or other useful information
6	**Morphosyntactic information**			

	a	Part-of-speech marker	part of speech of the headword (or the secondary headword)	code	helps both types of user find the information they are looking for, by focussing the search
	b	Inflectional class	Inflectional paradigm of the entry	code	helps SL user use TL item correctly helps TL user disambiguate TL word helps TL user use SL item correctly helps SL user disambiguate SL word
	c	Derivation	Cross-part-of-speech-information, morphologically derived forms	SL	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	d	Gender	Information about the gender of the entry in SL and TL	code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	e	Number	Information about the grammatical number of the entry in SL and TL	code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	f	Mass vs. Count	Information whether a noun is mass or count, in SL and TL	code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	g	Gradation	For adverbs and adjectives	code	helps SL user use TL item correctly helps TL user disambiguate TL word
7	Subdivision counter		indicates the start of new section or subsection (*'sense'*)	number / letter	'signpost' helping user to find their way about the entry more efficiently

8	Entry subdivision		separate section or subsection in entry (often called *dictionary sense*)	Dictionary text	breaks up entry, making it easier to read and find what is being sought
9	Sense indicator		synonym or paraphrase of headword in this sense, or other brief sense clue indicating specific sense of SL or TL item	SL	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
10	Linguistic label		the style, register, regional variety, etc. of the SL or TL item	code	helps SL user identify the sense of the headword helps both users translate helps TL user understand
11	**Syntactic information**				
	a	Subcategorization frame	(i.) Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.	code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	b	Obligatority of complements	Information whether a certain complement is obligatory or not	code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	c	Auxiliary	Which type of auxiliary is selected by a given predicate (in certain languages auxiliary selection is related to issues like unaccusativity, which on turn lies at the interface between lexicon and syntax)	code	acts as a sense indicator helps SL user select appropriate TL equivalent
	d	Light or support verb construction	Constructions with light verbs	SL or TL	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent

	e	Periphrastic constructions	Constructions containing periphrasis, usage, semantic value, etc.	SL or TL	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	f	Phrasal verbs	Particular representation of phrasal constructions	SL or TL	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	g	Collocator	(i.) typical subject /object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented etc.	SL or TL	acts as a sense indicator helps SL user select appropriateTL equivalent helps TL user translate or understand the SL item
	h	Alternations	Syntactic alternations an entry can enter into	Code	acts as a sense indicator
12	**Semantic information**				
	a	Semantic type	Reference to an ontology of types which are used to classify word senses	Code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	b	Argument structure	Argument frames, plus semantic information identifying the type of the arguments, selectional constraints, etc.	Code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	c	Semantic relations	Different types of relations (e.g. synonymy, antonymy, meronymy, hyperonymy, Qualia Roles, etc.) between word senses, etc.	Code	acts as SL sense indicator for SL user acts as TL sense indicator for TL user

	d	Regular polysemy	Representation of regular polysemous alternations	Code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	e	Domain	Information concerning the terminological domain to which a given sense belongs	Code	helps SL user identify the sense of the headword or other SL item helps TL user identify the sense of a TL equivalent
	f	Decomposition	Representation of relevant meaning component, e.g. causativity, agentivity, motion, etc.	Code	acts as SL sense indicator for SL user acts as TL sense indicator for TL user
13	Translation		TL equivalent of SL item	TL	helps TL user understand helps both users translate
14	Gloss		TL explanation of meaning of an SL item which has no direct equivalent in the TL	TL	helps TL user understand helps both users translate
15	Near-equivalent		TL item corresponding to an SL item which has no direct equivalent in the TL	TL	helps TL user understand helps both users translate
16	Example phrase (straightforward)		a phrase or sentence illustrating the non-idiomatic use of the headword, in a context where the TL equivalent is virtually a word-to-word translation	SL	acts as SL sense indicator for SL user acts as TL sense indicator for TL user helps TL & SL users to use the foreign-language item correctly
17	Example phrase (problematic)		a phrase or sentence illustrating a non-idiomatic use of headword in a context where a specific TL equivalent is required (*i.e. an SL example which is easily understandable for the TL speaker, but presents translation problems for the SL speaker*)	SL	helps SL user avoid a translating error acts as a sense indicator for SL user helps TL user subsequently to use the SL item correctly

18	Multiword unit	(idiomatic) multiword expression (MWE) containing the headword *(the term MWE covers idioms, fixed & semi-fixed collocations, compounds etc*.)	SL	helps both users translate
19	Subheadword *also* secondary headword	lemma morphologically related to the headword, figuring as head of a sub-entry (*subheadwords can be compounds, phrasal verbs, etc.)*	SL	saves space helps both types of user find the information they are looking for
20	Usage note	how the headword is used; 'macro' information which cannot appear at every appropriate entry; warning of cultural differences between the two languages; etc.	SL *or* TL	helps both types of user to avoid misunderstandings about the foreign language item, based on own-language knowledge
21	Frequency	Information about the frequency of the entry	code	helps both users translate

3 Survey of relevant representative lexicons

In order to better analyze lexicons, we organized the present 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.

3.1 MRDs

3.1.1 Collins, Collins Gem, Hachette-Oxford, Oxford dictionaries and the dictionaries browser DicoPro

3.1.1.1 Survey of the Dictionaries

- **Collins**

Collins Italian/English - English/Italian Dictionary

Languages: English - Italian, Italian - English

Published: 5/11/95

Official Description: Over 160,000 references and 230,000 translations

Queriable: via the ***public*** **DicoPro** Browser (*http://dicopro.unige.ch/DicoProPublic/*, see section 3.1.1.2)

- **Collins Gem**

Collins Gem - French Dictionary [Fourth edition]

Languages: French - English

Published: 3/1/97

Official Description: Over 40,000 references and 70,000 translations; extensive coverage of current French and English; clear, attractive typography for quick and easy access; special entries on French life and culture

Queriable: via the ***internal*** **DicoPro** Browser

Collins Gem - German Dictionary [Fourth edition]

Languages: German - English

Category: School and college

Published: 3/1/97

Official Description: Over 40,000 references and 70,000 translations; extensive coverage of current German and English; clear, attractive typography for quick and easy access; special entries on German life and culture; contains details of German spelling reform

Queriable: via the ***internal*** **DicoPro** Browser

- **Hachette-Oxford**

Hachette-Oxford/Oxford-Hachette French Dictionary

Languages: English - French

Published: April 1997

Official Description: Thousands of example sentences, taken from real speech and written sources; guide the user; over 350,000 words and phrases, and over 530,000 translations provide the most comprehensive and up-to-date coverage of the general, scientific, literary, and technical vocabulary of contemporary French and English;

historical, idiomatic, colloquial, and regional French are also generously covered, etc.

Queriable: via the **internal DicoPro Browser** (see below)

- **Oxford**

Oxford Spanish-English Dictionary [Second edition]

Languages: Spanish - English

Published: 19-02-1998

Official description: For this second edition, the Oxford Spanish Dictionary has been extensively revised and updated; new features include comprehensive language notes within the text, in addition to new boxed notes giving information on subject areas such as games and sports, colours, the human body and time.

Queriable: via the **internal DicoPro Browser** (see below)

3.1.1.2 Browser: DicoPro (http://dicopro.unige.ch/DicoProPublic/)

The browser *DicoPro* has been developed in the project DicoPro (*On-line Dictionary Consultation for Language Professionals on intranet*), a project funded within the Multilingual Information Society Programme (MLIS). The project was funded by the European Union and the Swiss Federal Office of Science and Education. (For a full list of partners, detailed project reports and an on-line demo cf. *http://www.issco.unige.ch*.) The aim of the project was to develop a uniform, platform-independent interface for accessing multiple dictionaries and other lexical resources via the Internet/intranets. The project brought together technical experts for program development, major dictionary publishers providing data and insight into usage of the data and language professionals for testing and validation of the tool.

The background to this project was a dictionary server (DICO) with similar functionalities, but running on a local area network. The DICO system, developed in 1990, was based on a client-server architecture and offered two interfaces xdico and tdico, to accommodate Unix workstations running X-Windows and PCs via a simple terminal mode. The program has been operational on the University of Geneva network. It provides access to ten mono- and bilingual dictionaries and is still regularly consulted by hundreds of users. The MLIS DicoPro project can thus be seen as a natural next generation of dictionary servers, taking full advantage of the Internet and the growing potential of e-commerce.

The DicoPro consortium developed what is anticipated will be a commercially viable tool based on existing open standards. The data formats used in the system rely on SGML, HTML and XML technologies. The client and server tools have been developed to run on a wide range of platforms. In particular, all development was done using the portable programming language Java. In this section, we describe the core components of the system in somewhat more detail.

- The dictionary data

A number of bilingual and monolingual dictionaries were supplied by the DicoPro consortium partners for use in the project. Typically, source data obtained from project partners was marked up in SGML-like fashion.

- Converting the data : XMLTrans

To transform dictionary SGML-like entries for display in HTML, a transformation tool, XMLTrans, was developed for DicoPro. For each dictionary, a set of XMLTrans transformation rules was written and then iteratively improved them until the resultant HTML was satisfactory.

XMLTrans was also used to extract relevant fields from entries for indexing. For instance, the translation component of a bilingual entry can be extracted and indexed to allow the user to search the dictionary using only the translation fields of entries.

- The DicoPro server

Once prepared, data is stored on the DicoPro server, which is a robust cross platform Java program. It was developed using a threaded design, allowing it to handle many concurrent users accessing diverse data. The server can be run as either a standalone application, or as a Servlet from within a web server such as Apache. This second model permits filtering of clients by IP address and the use of SSL encryption.

- The DicoPro client

The client is also a cross platform Java application which can be run on Windows, Unix or Macintosh systems. An applet version of the client runs from within a web browser.

The client connects to the Dictionary Information Server (DIS) which provides it with a list of available dictionaries (fig.1). Once opened, each dictionary has its own space with its own menus and options for searching and displaying results. Multiple dictionaries can be opened and consulted at the same time (fig. 2). A number of indexes such as prefix, suffix, regular expression, and inflected form are available.

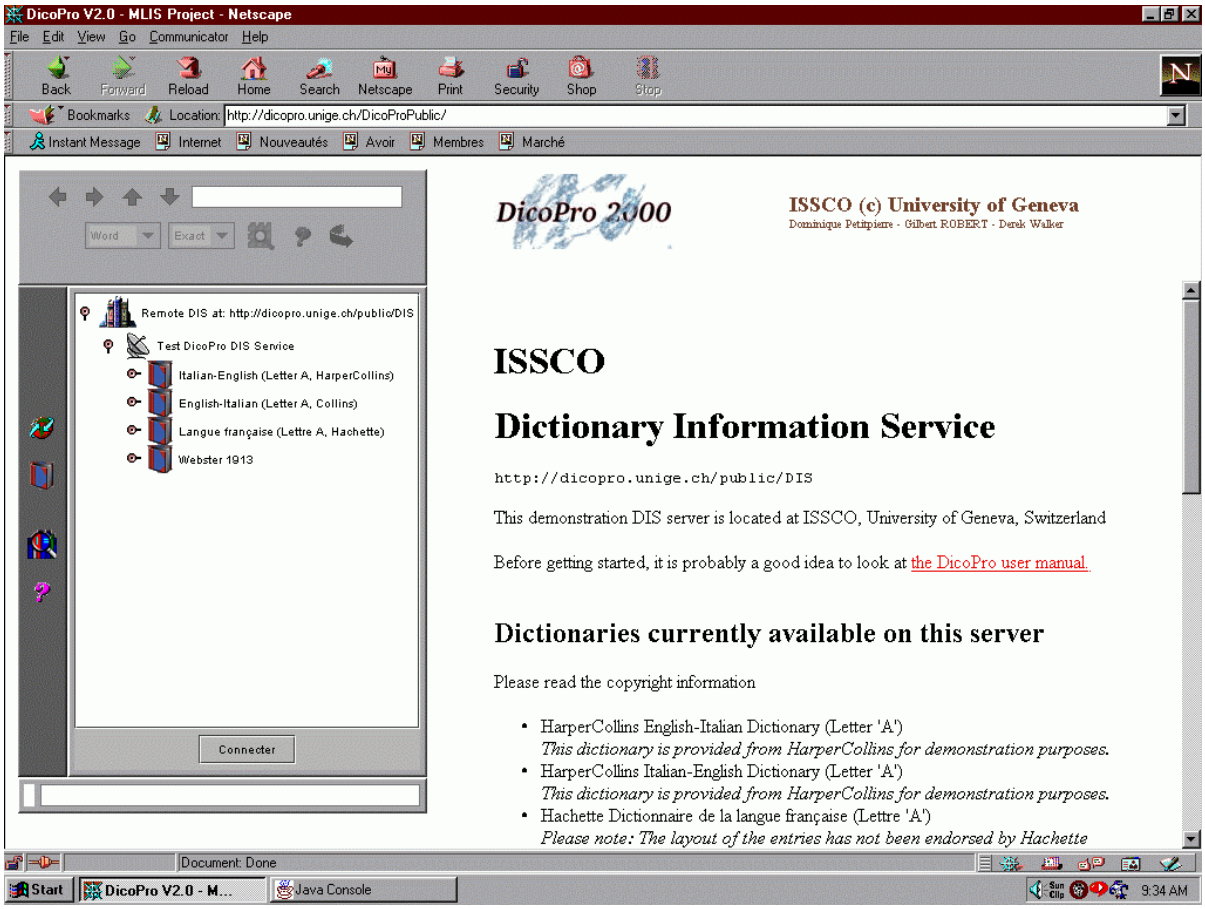


Fig. 1: DicoPro Browser

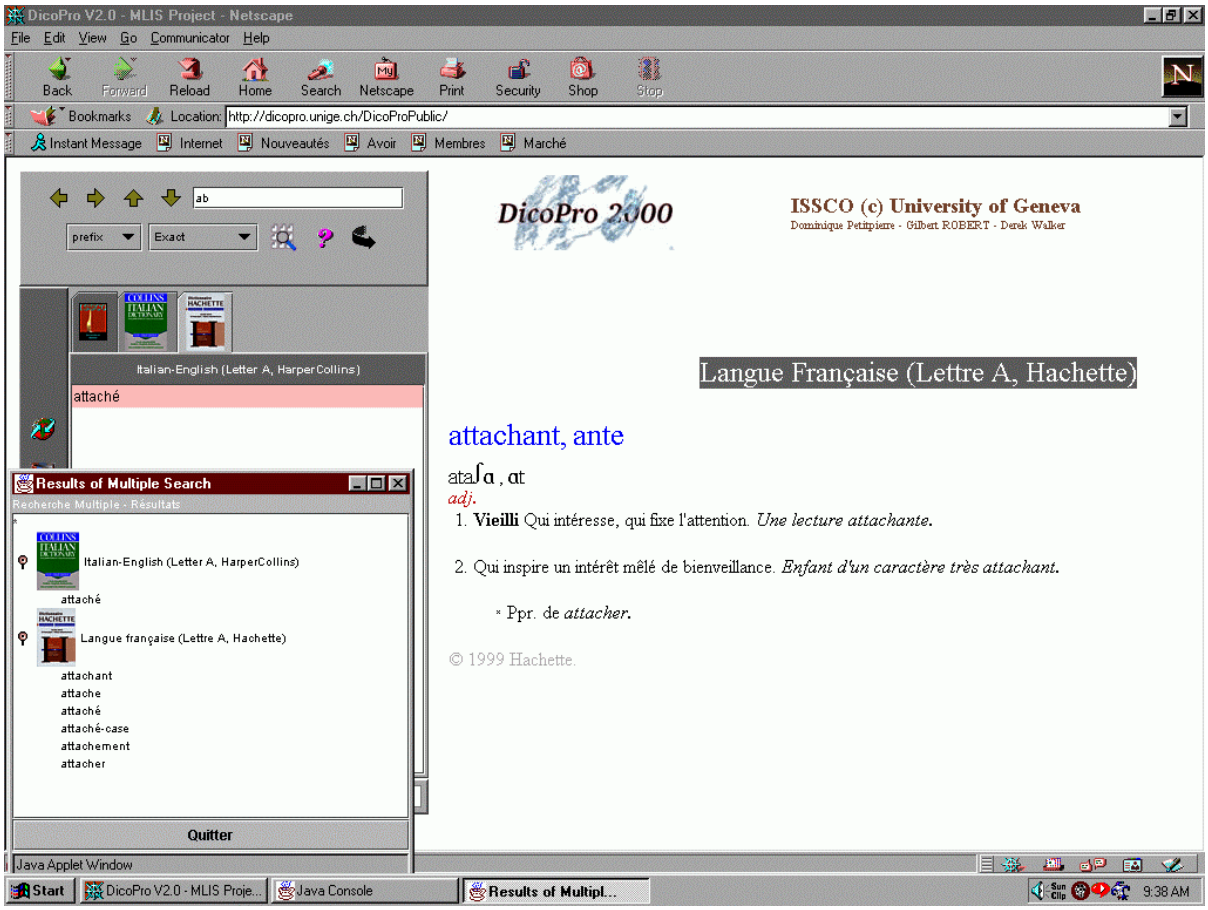


Fig. 2: Accessing multiple lexical resources

The client software enables simultaneous access to multiple lexical resources from a diversity of well-respected publishers. It provides a uniform interface allowing parallel queries in multiple dictionaries, regardless of the actual physical location of the resource. Each user (or user group) can select the set of dictionaries to be consulted (fig. 2).

3.1.1.3 Synoptic tables of information types in the dictionaries

In the following we give an overview of the content of the dictionaries investigated in this survey on the basis of the table “Lexical Information in Bilingual Resources” (see chapter 2). Our aim is to verify if this information is available and if it can be easily extracted. The idea is thus to see whether the encoding of the XML tags corresponds to the organisation of a typical dictionary entry and to gather differences in the organisation. It is important to mention that we did not have any user manual or any other explanatory documents describing the XML tags, which implies that we had to figure out on the basis of examples what a particular XML tag was supposed to encode exactly.

This section presents the result of this work for the 4 different types of dictionaries presented in section 3.1.1.1, i.e. *Collins*, *Collins Gem*, *Oxford-Hachette* and *Oxford*. As mentioned the work was carried out on the basis of the source format of the dictionaries, but for the examples of this report, the tags were anonymised. Instead we have integrated an entry for each dictionary as it appears through the DicoPro Browser (see section 3.1.1.2). Each table contains the following type of information:

- *entry component* (according to the proposed table)

- *corresponding XML tag(s)* in the source format of the dictionary in question.
  Possible values:
  none: there is no XML tag corresponding to the entry component.
  one: there is exactly one XML tag corresponding to the entry component.
  one (+info number-of-entry-component): there is one XML tag corresponding to the entry component. Moreover, the field contains information concerning another entry component indicated in *number-of-entry-component*. More details are given in the column " *comments* ".
  several: there are several XML tags corresponding to the entry component, i.e. the dictionary entry as it is organised makes a more careful distinction of the information gathered in the entry component in question.
  several (+info number-of-entry-component): there are several XML tags corresponding to the entry component. Moreover, the corresponding XML tags contain information concerning another entry component. More details are given in the column " *comments* ".
  common tag (number-of-entry-component): there is no XML tag corresponding exactly to the entry component in question, but the information is gathered under the entry component which is specified in *number-of-entry-component*. Therefore this other entry component is described as either *one (+info number-of-entry-component)* or *several (+info number-of-entry-component)*. More details are given in the column " *comments* ".
  (SL): additional specification which indicates whether it concerns the source language (SL) where this is necessary.
  (TL): additional specification which indicates whether it concerns the target language (TL) where this is necessary.

- *comments*: any comments concerning the relation between the XML tag(s) and the entry component or the type of information in question.

3.1.1.3.1 Collins

Table 2: Collins: Table comparing entry components and XML tags

	Entry component		Present	Corresponding XML tags	Comments
1	Headword		✓	several	different tags distinguish acronyms, compounds, etc.
2	Phonetic transcription		✓	one	
3	Variant form		✓	one	e.g. "coloured" (headword) - "colored" (variant form)
4	Inflected form		✓	one (SL) & one (TL)	
5	Cross-reference		✓	one	to (another) headword
6	**Morphosyntactic information**				
	a	Pos marker	✓	one (+info 6de & 11a)	can include information concerning the number and gender (entry component 6de), e.g. "noun sg", "noun pl" and the subcatgorization (entry component 11a), e.g. "transitive verb"
	b	Inflectional class			
	c	Derivation			
	d	Gender	✓	common tag (6a) & one (TL)	The information concerning the SL is classified in the Pos marker (entry component 6a).
	e	Number	✓	common tag (6a) & one (TL)	The information concerning the SL is classified in the Pos marker (entry component 6a).
	f	Mass vs. count			
	g	Gradation	✓	one	
7	Subdivision counter		✓	one	mode: number
8	Entry subdivision		✓	one	mode: letter
9	Sense indicator		✓	several	domain (e.g. "Music", "Biology") and semantic information (e.g. "person", "degree", etc.)
10	Linguistic label		✓	several (+info 12e, 20)	different tags for region (e.g. "Am", "Brit"), register (e.g. "familiar", "literaryhistoric co"), historic context (e.g. "Old"), usage (e.g. "fig"), etc. Might contain information about usage (entry component 20)
11	**Syntactic information**				
	a	Subcatego-rization frame	✓	common tag (6a & 16)	there seems not to be a tag which encodes the structure as such. The information is contained sometimes in the Pos marker (entry component 6a), or must be extracted from the Example phrase (entry component 16 & 17).
	b	Obligatority of complements			
	c	auxiliary			
	d	Light or support verb	✓	common tag (16)	This information must be extracted from Example phrase (entry component 16). It is

		construction			not made explicit. It might occur as a headword (entry component 1).
	e	Periphrastic construction	✓	common tag (16)	This information must be extracted from Example phrase (entry component 16). It is not made explicit. It might occur as a headword (entry component 1).
	f	Phrasal verbs	✓	common tag (16)	This information must be extracted from Example phrase (entry component 16). It is not made explicit. It might occur as a headword (entry component 1).
	g	Collocator	✓	common tag (16)	This information must be extracted from Example phrase (entry component 16). It is not made explicit. It might occur as a headword (entry component 1).
	h	Alternation			
12	**Semantic information**				
	a	Semantic type			
	b	Argument structure			
	c	Semantic relation	✓	common tag (5)	
	d	Regular polysemy			
	e	Domain	✓	common tag (10)	
	f	Decomposition			
13	Translation		✓	several (+info 14 & 15)	several tags are used for the translation to distinguish an acronym, its expansion, collocations, etc. There seems not to be a direct relationship between the tags and the distinction proposed here (translation, gloss, Near-equivalent).
14	Gloss		✓	common tag (13)	there seems not to be a direct relationship between the tags and the distinction proposed here (translation, gloss, Near-equivalent).
15	Near-equivalent		✓	common tag (13)	there seems not to be a direct relationship between the tags and the distinction proposed here. There is however a tag which seem to correspond to what could be called an "approximate translation".
16	Example phrase (straightforward)		✓	one (+info 11adefg, 12, 17 & 18)	the entry components 16-18 seem to be basically all correspond to one tag.
17	Example phrase (problematic)		✓	common tag (16)	is classified together with the general Example phrase (entry component 16)
18	multiword unit		✓	common tag (16)	is classified together with the general Example phrase (entry component 16). It might occur as a headword (entry component 1).
19	Subheadword (secondary headword)				might appear under the general Example phrase (entry component 16), evt. the headword (entry component 1) itself, given that there are several tags for this entry component. Depends what is exactly meant by subheadword.
20	Usage note		✓	common tag (10)	has been classified under linguistic label (entry component 10).
21	Frequency				

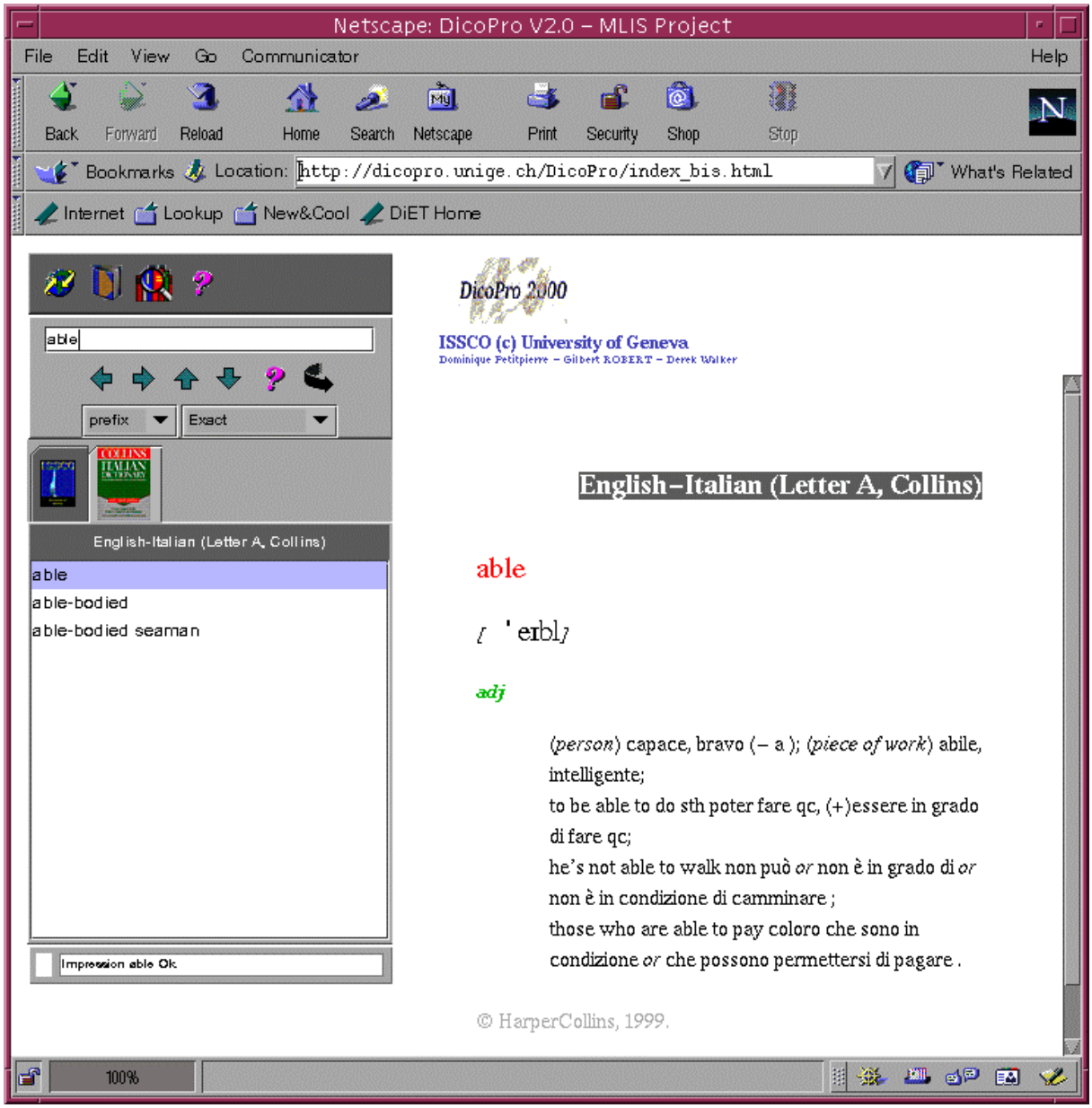


Fig. 3: example of the entry "able" (DicoPro Browser)

3.1.1.3.2 Gem

Table 3: Gem: Table Comparing entry components and XML tags

	Entry component	Present	Corresponding XML tags	Comments
1	Headword	✓	several (+info 11d-g, 16-19)	In the Collins Gem, basically every example phrase, idiom, subheadword, etc. is treated as separate entry (e.g. "à" (at), "à trois heures" (at three o'clock), "à bicyclette" (by bicycle) are three separate entries). Therefore the headword contains information which actually correspond to the entry component 16-19. But there is a tag that distinguishes the main headword from the related entries. E.g. for the entry "avant, à l'avant" (in front), "avant" is tagged by means of the main headword tag, and "à l'avant" is tagged by means of the secondary headword tag.
2	Phonetic transcription			
3	Variant form	✓	one (+info 4)	e.g. "clé"(key) (headword) - "clef" (variant form). The same tag is also used to encode information about the inflected form (entry component 4).
4	Inflected form	✓	common tag (3)	is classified together with the variant form (entry component 3).
5	Cross-reference	✓	several (+info 12e)	two tags depending whether (i) the headword (entry component 1) corresponds to the abbreviation of the full word encoded here or whether (ii) it is a cross-reference to a synonym, hyperonym, etc.
6	**Morphosyntactic information**			
	a Pos marker	✓	one (+info 6de & 11a)	can include information concerning the gender or the number (entry component 6de), e.g. "noun sg", "noun pl" and the subcategorization frame (entry component 11a), e.g. "transitive verb".
	b Inflectional class			
	c Derivation			
	d Gender	✓	common tag (6a) (SL) & one (+info 6e) (TL)	the information concerning the SL is classified together with the Pos marker (entry component 6a).
	e Number	✓	common tag (6a) (SL) & common tag (6f) (TL)	the information concerning the SL is classified together with the Pos marker (entry component 6a).
	f Mass vs. count			
	g Gradation			
7	Subdivision counter	✓	one	
8	Entry subdivision	✓	one	
9	Sense indicator	✓	one (+ info 10)	specifies information such as “ animal ”, “ house ”, “ direction ”, but also “ fig ” etc. (see entry component 10)

10	Linguistic label	✓	common tag (9)	e.g. “ fig ”, etc.	
11	**Syntactic information**				
	a	Subcategorization frame	✓	common tag (1 & 6a)	there seem not to be a tag which encodes the subcategorization as such. The information is contained sometimes in the Pos marker (entry component 6a), sometimes directly in the headword (entry component 1).
	b	Obligatority of the complements			
	c	Auxiliary			
	d	Light or support verb	✓	common tag (1)	usually a separate entry (i.e. tag as headword (entry component 1)
	e	Periphrastic constructions	✓	common tag (1)	usually a separate entry (i.e. tag as headword (entry component 1)
	f	Phrasal verbs	✓	common tag (1)	usually a separate entry (i.e. tag as headword (entry component 1)
	g	Collocator	✓	common tag (1)	usually a separate entry (i.e. tagged as headword (entry component 1)
	h	Alternations			
12	**Semantic information**				
	a	Semantic type	✓	common tag (9)	information such as personne, animal, etc.
	b	Argument structure	✓	common tag (9)	information such as suj:personne
	c	Semantic relations	✓	common tag (5)	
	d	Regular polysemy			
	e	Domain	✓	common tag (9)	
	f	Decomposition			
13	Translation		✓	one (+info 14-15)	one tag is used for the translation, gloss and Near-equivalent, i.e. for the entry components 13-16.
14	Gloss		✓	common tag (13)	one tag is used for the translation, gloss and Near-equivalent, i.e. for the entry components 13-16.
15	Near-equivalent		✓	common tag (13)	one tag is used for the translation, gloss and Near-equivalent, i.e. for the entry components 13-16.
16	Example phrase (straightforward)		✓	common tag (1)	usually a separate entry (i.e. tagged as headword (entry component 1)
17	Example phrase (problematic)		✓	common tag (1)	usually a separate entry (i.e. tagged as headword (entry component 1)
18	Multiword unit		✓	common tag (1)	usually a separate entry (i.e. tagged as headword (entry component 1)
19	Subheadword (secondary headword)		✓	common tag (1)	usually a separate entry (i.e. tagged as headword (entry component 1)
20	Usage note		✓	common tag(10)	same tag as for linguistic label (entry component 10)
21	Frequency				

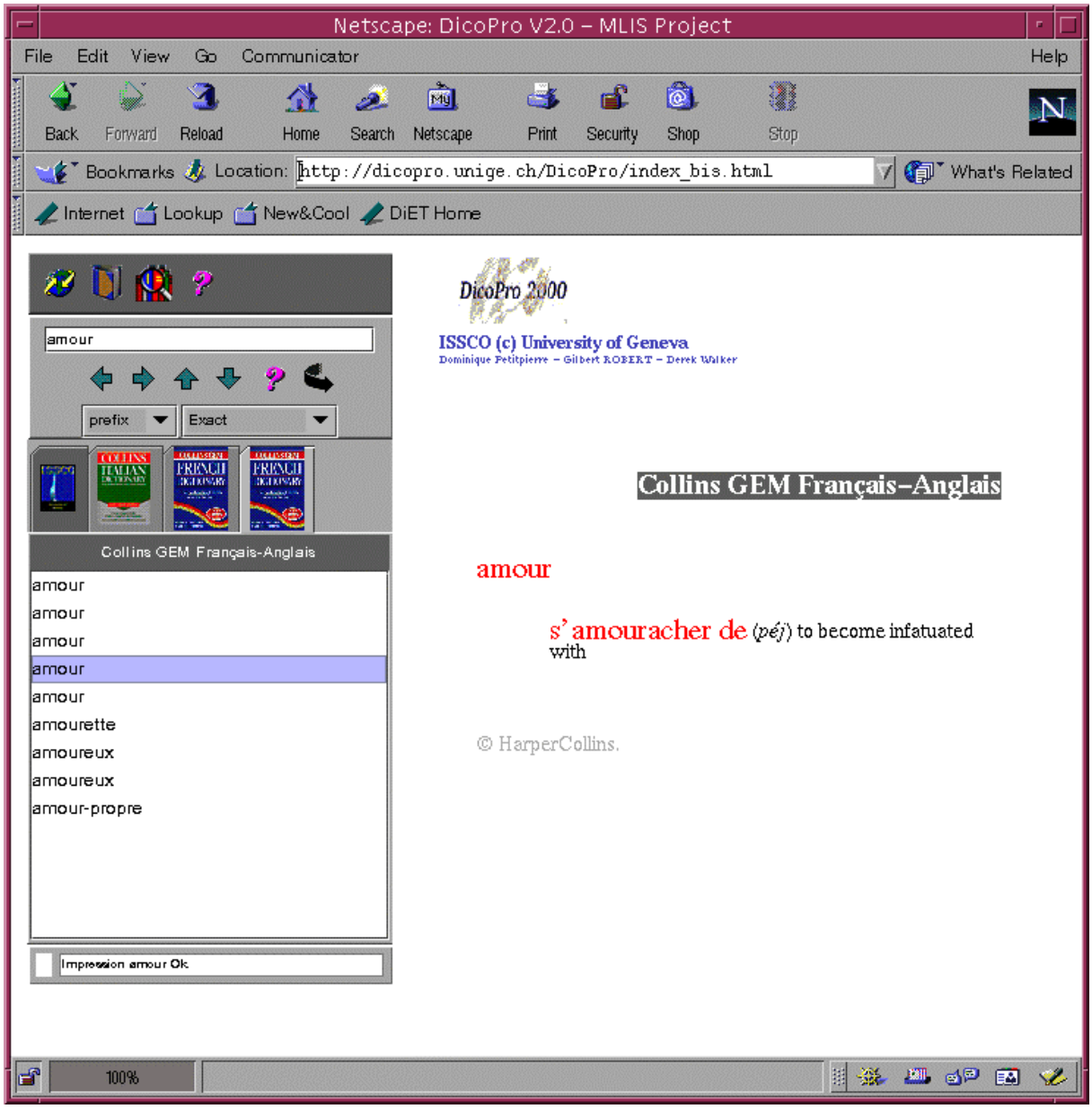


Fig. 4: Example of the entry "amour" (DicoPro Browser)

3.1.1.3.3 Oxford Hachette

Table 4: Oxford-Hachette: Table comparing entry components and XML tags

	Entry component		Present	Corresponding XML tags	Comments
1	Headword		✓	several (+info 18)	depending on whether the headword is a "compound entry" , a "standard entry", or a "no-root entry", there are different basic entry tags. Furthermore the tag referring to the headword is different, depending on the type of entry. If it is a standard entry, the headword is described by one tag. If the entry is a "compound word", there are two other tags for describing it: one tag for the compound (e.g. "accession number") and one tag for the base word (e.g. " accession ").
2	Phonetic transcription		✓	one	+ hierarchical tag which contains phonetics and related label.
3	Variant form		✓	one	
4	Inflected form		✓	one	
5	Cross-reference		✓	several	different tags corresponding to 'global' cross-references (which relate to the entire entry), to a sense number in a cross reference, to a verb table reference or to a target word cross reference.
6	**Syntactic information**				
	a	Pos marker	✓	one (+info 6de & 11a)	includes information about the structure (entry component 11a), e.g. "transitive verb" or the number and gender (entry component 6de).
	b	Inflectional class			
	c	Derivation			
	d	Gender	✓	common tag (6a) (SL) & one (+info 6e) (TL)	
	e	Number	✓	common tag (6a) & common tag (6d) (TL)	
	f	Mass vs count			
	g	Gradation			
7	Subdivision counter		✓	several	
8	Entry subdivision		✓	one	
9	Sense indicator		✓	several (+info 10 & 11e)	different tags: global usage ("instruments", "professions", etc.), domain label tagged as linguistic label (entry component 10).
10	Linguistic label		✓	common tag (9 + 20)	one tag to describe domain(e.g. "archeology"), register, nationality, etc.
11	**Syntactic information**				
	a	Subcategorization frame	✓	several (+info 11d,g & 16)	there are different tags to give information about the structure, for example for preposition groups, phrasal verb patterns, fixed and semi-fixed patterns, also separate tags for idioms. Not always quite clear to distinguish from example (entry component 16).

	b	Obligatority of complements			
	c	Auxiliary			
	d	Light or support verb constructions	✓	one (+info 16)	
	e	Periphrastic constructions	✓	one (+info 16)	
	f	Phrasal verbs	✓	one (+ info 16)	
	g	Collocator	✓	one (+info 16)	
	h	Alternations		none	
12	**Semantic information**				
	a	Semantic type			
	b	Argument structure			
	c	Semantic relations			
	d	Regular polysemy			
	e	Domain	✓	common tag (9)	
	f	Decomposition			
13	Translation		✓	common tag (1)	see headword (entry component 1)
14	Gloss		✓	one	
15	Near-equivalent				
16	Example phrase (straightforward)		✓	common tag (11a)	sometimes difficult to distinguish from subcategorization frame (entry component 11a).
17	Example phrase (problematic)		✓	common tag (16)	is classified together with the general Example phrase (entry component 16).
18	Multiword unit		✓	common tag (1)	see headword (entry component 1)
19	Subheadword (secondary headword)		✓	one	see headword (entry component 1)
20	Usage note		✓	common tag (10)	is classified under linguistic label (entry component 10).
21	Frequency				

Example of an entry

<tag1><tag2>able</tag2> <tag2>< tag4>"eIbl</tag4></tag3>
<tag5><tag6>adj</tag6></tag5> <tag7><tag8>to be able to</tag8> meaning <tag8>can</tag8> is usually translated by the verb <tag8>pouvoir</tag8>: <tag8>I was not able to go</tag8> = je ne pouvais pas y aller; <tag8>I was not able to help him</tag8> = je ne pouvais pas l'aider. The main exception to this occurs when <tag8>to be able to</tag8> implies the acquiring of a skill, when <tag8>savoir</tag8> is used: <tag8>he's nine and he's still not able to read</tag8> = il a neuf ans et il ne sait toujours pas lire.<tag9>For more examples and other uses, see the entry below.</tag9></tag7> <tag10>(<tag11>having ability to</tag11>) <tag12>to be &hw. to do/be</tag12> pouvoir faire/&ec.tre; <tag13>he was/wasn't &hw. to read it</tag13> il pouvait/ne pouvait pas le lire; <tag13>she was &hw. to play the piano at the age of four</tag13> elle savait jouer du piano &ag. quatre ans; <tag13>I'll be (better) &hw. to give you more information after the meeting</tag13> je serai en mesure de <tag14>or</tag14> je pourrai vous donner plus de renseignements apr&eg.s la r&ea.union</tag10>; <tag15>(<tag11>skilled</tag11>) <tag16>lawyer, teacher etc</tag16> comp&ea.tent; (<tag11>gifted</ tag11>) <tag16>child</tag16> dou&ea.</tag15>.</tag1>

3.1.1.3.4 Oxford

Table 5: Oxford: comparing entry components and XML tags

	Entry component	Present	Corresponding XML tags	Comments
1	Headword	✓	several	depending on whether the headword is an initial-letter headwork (e.g. "A") or not. Sometimes there can be several entries for the same headword - in those cases there is an additional tag which enumerates them.
2	Phonetic transcription	✓	one	
3	Variant form	✓	several	there are many tags which describe the variant form: the variant form itself, the feminine variant form, the plural variant form, the part-of-speech of the variant form, the regional label of the variant form, the expansion if it is an abbreviation, etc. The variant form is thus treated like a 'separate entry' in the principal entry.
4	Inflected form	✓	several	different tags depending on whether the inflected form is feminine, plural, etc.
5	Cross-reference	✓	several	different tags corresponding to 'global' cross-references, reference to a proverb, to the expansion of an abbreviation, etc.
6	**Morphosyntactic information**			
	A Pos marker	✓	one (+info 6de & 11a)	can include information concerning the grammar marker (entry component 6de), e.g. "noun sg", "noun pl" and the structure (entry component 11a), e.g. "transitive verb".
	B Inflectional class			
	C Derivation			
	D Gender	✓	common tag (6a)	
	E Number	✓	common tag (6a)	
	F Mass vs count			
	G Gradation			
7	Subdivision counter	✓	several (+info 8)	depending on the number of senses. Furthermore there are "hierarchical" tags which define the "part-of-speech section", the "compound section" and the "verb section".
8	Entry subdivision	✓	common tag (7)	difficult to distinguish between "subdivision counter" and "entry subdivision" See entry component 7.
9	Sense indicator	✓	several	they exist both for the SL and the TL.
10	Linguistic label	✓	several	serveral tags depending on the register, style, region, etc. They exist both for the SL and the TL.
11	**Syntactic information**			
	A subcategorization frame	✓	several	there are different tags to give information about the structure, a general one, plus specific ones (pronominal verb, infinitive constructions, etc.) They exist both for the SL and the TL (e.g. complementation pattern of the translation).

	B	Obligatority of the complement			
	c	Auxiliary			
	d	Light or support verb construction	✓	common tag (16)	different tags corresponding to the collocate of the adjective, the adverb, the subject, the object or the prepositional collocate of the verb.
	e	Periphrastic constructions	✓	common tag (16)	
	f	Phrasal verbs	✓	common tag (16)	
	g	Collocator	✓	several	different tags corresponding to the collocate of the adjective, the adverb, the subject, the object or the prepositional collocate of the verb.
	h	Alternations			
12	**Semantic information**				
	a	Semantic type			
	b	Semantic relations			
	d	Regular polysemy			
	e	Domain	✓	common tag (9)	
	f	Decomposition			
13	Translation		✓	several	there are different tags corresponding to the:'standard' translation, translation of an abbreviation, translation of an example, feminine form of translation.
14	Gloss		✓	one	
15	Near-equivalent		✓	several	there are different tags corresponding to the cultural equivalent, a definition (if not translation possible), encyclopaedic information to the translation, translation of an idiom, translation of a contextualized example of a verb compound, translation of a proverb.
16	Example phrase (straightforward)		✓	several (+ info 17)	there are different tags corresponding to the 'standard' example, examples of a diminutive form, examples of an idiom or of a proverb, contextualized examples of an idiom or of a proverb, examples in a note, etc.
17	Example phrase (problematic)		✓	common tag (16)	See entry component 16.
18	Multiword unit		✓	several	there seems to be a specific label for verb compounds.
19	Subheadword (secondary headword)				See entry component 1.
20	Usage note		✓	several	is classified under linguistic label (entry component 10).

Example of an entry

<tag1><tag2>abarcar</tag2> <tag3>A2</tag3> <tag4>vt</tag4> <tag5> <tag6>temas/materias</tag6> <tag7>to cover</tag7>; <tag8>el programa abarca desde la Reconquista hasta el siglo XIX</tag8> <tag9>the program takes in <tag10>o</tag10> covers <tag10>o</tag10> spans the period from the Reconquest to the 19th century</tag9>; <tag8>sus tierras abarcan desde el r&ia.o hasta la sierra</tag8> <tag9>his land stretches <tag10>o</tag10> extends from the river up to the mountains</tag9>; <tag8>abarcaba todo el territorio que ahora se conoce como Uruguay</tag8> <tag9>it extended over <tag10>o</tag10> embraced <tag10>o</tag10> spanned <tag10>o</tag10> included all the territory now known as Uruguay</tag9></tags4> <tags11> (<tag12>dar abasto con</tag12>) <tag13>trabajos/actividades</tag13> <tag7>to cope with</tag7>; <tag8>se ha echado encima m&aa.s de lo que puede &swing.</tag8> <tag9>he's bitten off more than he can chew</tag9>, <tag9>he's taken on more than he can cope with</tag9>; <tag14>quien mucho abarca poco aprieta</tag14> <tag15>don't try to take on too much (<tag10>o</tag10> you've/he's taken on too much <tag10>etc</tag10>)</tag15></tag11> <tag16> (<tag12>con los brazos</tag12>) <tag7>to embrace</tag7>, <tag7>encircle</tag7>; <tag8>no le abarco la mu&nt.eca con la mano</tag8> <tag9>I can't get my hand around his wrist</tag9></tags4> <tags4 let=d> (<tag12>con la mirada</tag12>) <tag7>to take in</tag7></tag16></tag1>

Comments about the structure of the dictionaries : a lot of information is encoded in printed dictionaries, but it is encoded very differently, even from one language pair to another. The ideal representation of a typical bilingual dictionary entry is not always the one followed by the dictionaries. Information is contained but not made explicit by the XML tags. For example, *Sense indicator*, *Linguistic label* and *usage notes* are mostly stored in the same field. Similarly, *Problematic* and *Straightforward example phrases* are not distinguished.

3.1.2 Multilingual information in the Van Dale lexicons

3.1.2.1 Description

The Van Dale bilingual dictionaries are developed for native speakers of Dutch. This means that the resources contain only very limited information on the Dutch words and much more information on the foreign-language target words. We here give a description of the Dutch-English and English-Dutch dictionaries (Martin and Tops, 1986) but the other dictionaries have similar structures and content.

The dictionaries are available on electronic tapes from which the printed books have been derived. The information is stored in separate fields with field-names and values. Some values are restricted to codes, others contain free text. The entry-structure is homograph-based but homographs are distinguished only when the part-of-speech differs and/or the pronunciation. Sub-homographs are used when senses differ in major grammatical properties such as valency, countability, predicate/attributive usage.

Two types of translations are given:

- main translation: more general, always applicable
- secondary translation: more specific, often limited to some contexts or constraint

A main translation is always present. Secondary translations are optional, and are often limited either stylistically or semantically (e.g. verbal selectional restrictions). Still, the secondary translations are often better translations than the main translations.

Table 6: Number of entries, senses and translations in the Van Dale

Dutch-English & English-Dutch dictionaries

	Dutch-English	English-Dutch
Entries	90,925	89,428
Senses	127,024	156,838
Main Translations	145,511	152,318
Secondary Translations	104,181	162,752

The morpho-syntactic information is limited. In addition to POS, there are codes for countability, valency, plural/singular forms. A special system is used for the examples. In each example, the entry word is combined with a typical example word that is marked. The POS of the combination is indicated in the example number. For each sense of an entry, there will likewise be examples in which it is combined with a preposition, noun, verb, adjectives, etc., if such usage is typical for the word in that meaning. Figurative usage is also marked. Examples are tranlated and these example translations can have various codes and labels.

In addition to the grammatical and example information on the words and the translations, the dictionary contains a large amount of semantic information restricting the senses and the translations. In the case of the Dutch-English dictionary, we find for example the following additional information:

- [Sense-indicators] (53368 tokens) to specify the Dutch senses or polysemous entries. These contain bits and pieces from original definitions (often a genus word);
- [Biological gender marker] for English translations. This is necessary to differentiate translations when the source and target language have different words for male or female species: 286 translations are labelled as male, 407 translations as female;
- [Usage labels for domain, style and register] Applies to both Dutch senses and their English translations;
- [Dialect labels] for Dutch senses and their English translations;
- [Context markers] (23723 tokens, 16482 types). These are semantic constraints differentiating the context of multiple translations, and to limit the scope of translations having a narrower context than the Dutch source sense;

The usage labels and the domain labels are mostly stored in the same field. Differentiation has to be done by some parsing. The usage labels form a limited closed set of abbreviations and codes, the domain labels are free text. For the main-translations, about 400 different types of usage labels.

The translations can be single words, words combined with labels, co-ordination of translations and phrases. Phrasal translation may indicate a lexical gap in English or point to a multiword expression in the target language. Co-ordinations have been marked in the resource by "//" (for alternative words) or "/" (surrounding alternative phrases). This information can be used to split them in separate translation fields for a sense, e.g.:

gin//genever bottle

=> gin bottle; genever bottle

(administration of) the /last sacraments/extreme union/

=> administration of the extreme union; administration of the last sacraments; the last sacraments; the extreme union

(adult) literacy project//campaign

=> adult literacy project; adult literacy campaign; literacy project; literacy campaign

3.1.2.2 Synoptic tables of information types in the Van Dale lexicons.

Table 7: Lexical information in the Van Dale lexicon

	Entry component	Information content	Present
1	headword	lexical form(s) of the headword: how the headword is spelt	✓
2	Phonetic transcription	how the headword (or variant form etc.) is pronounced (in *International Phonetic Alphabet*)	
3	variant form	alternative spelling of headword or slight variation in the form of this word	✓
4	inflected form	other grammatical forms of the lemma (headword)	
5	Cross-reference	indication of another headword whose entry holds relevant	✓

			whose entry holds relevant information, or some other part of the dictionary where this may be found	
6	**Morphosyntactic information**			
	a	Part-of-speech marker	part of speech of the headword (or the secondary headword)	✓
	b	Inflectional class	Inflectional paradigm of the entry	
	c	Derivation	Cross-part-of-speech-information, morphologically derived forms	
	d	Gender	Information about the gender of the entry in SL and TL	✓
	e	Number	Information about the grammatical number of the entry in SL and TL	✓
	f	Mass vs. Count	Information whether a noun is mass or count, in SL and TL	
	g	Gradation	For adverbs and adjectives	
7	Subdivision counter		indicates the start of new section or subsection (*'sense'*)	✓
8	Entry subdivision		separate section or subsection in entry (often called *dictionary sense*)	✓
9	Sense indicator		synonym or paraphrase of headword in this sense, or other brief sense clue indicating specific sense of SL or TL item	✓
10	linguistic label		the style, register, domain, regional variety, etc. of the SL or TL item	✓
11	**Syntactic Information**			
	a	Subcategorization frame	(i.) Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.	
	b	Obligatority of complements	Information whether a certain complement is obligatory or not	
	c	Auxiliary	Which type of auxiliary is selected by a given predicate (in certain languages auxiliary selection is related to issues like unaccusativity, which on turn lies at the interface between lexicon and syntax)	

	d	Light or support verb construction	Constructions with light verbs	
	e	Periphrastic constructions	Constructions containing periphrasis, usage, semantic value, etc.	
	f	Phrasal verbs	Particular representation of phrasal constructions	✓
	g	Collocator	(i.) typical subject /object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented) etc.	✓
	h	Alternations	Syntactic alternations an entry can enter into	
12	**Semantic Information**			
	a	Semantic type	Reference to an ontology of types which are used to classify word senses	
	b	Argument structure	Argument frames, plus semantic information identifying the type of the arguments, selectional constraints, etc.	
	c	Semantic relations	Different types of relations (e.g. synonymy, antonymy, meronymy, hyperonymy, Qualia Roles, etc.) between word senses, etc.	
	d	Regular polysemy	Representation of regular polysemous alternations	
	e	Domain	Information concerning the terminological domain to which a given sense belongs	✓
	f	Decomposition	Representation of relevant meaning component, e.g. causativity, agentivity, motion, etc.	
13	Translation		TL equivalent of SL item	✓
14	Gloss		TL explanation of meaning of an SL item which has no direct equivalent in the TL	✓
15	Near-equivalent		TL item corresponding to an SL item which has no direct equivalent in the TL	✓
16	Example phrase (straightforward)		a phrase or sentence illustrating the non-idiomatic use of the headword, in a context where the TL equivalent is virtually a word-to-word translation	✓
17	Example phrase (problematic)		a phrase or sentence illustrating a non-idiomatic use of headword in	✓

	(problematic)	a context where a specific TL equivalent is required (*i.e. an SL example which is easily understandable for the TL speaker, but presents translation problems for the SL speaker*)	
18	multiword unit	(idiomatic) multiword expression (MWE) containing the headword *(the term MWE covers idioms, fixed & semi-fixed collocations, compounds etc*.)	✓
19	subheadword *also* secondary headword	lemma morphologically related to the headword, figuring as head of a sub-entry (*subheadwords can be compounds, phrasal verbs, etc.)*	
20	usage note	how the headword is used; 'macro' information which cannot appear at every appropriate entry; warning of cultural differences between the two languages; etc.	
21	Frequency	Information about the frequency of the entry	

Table 8: Linguistic Labels in the Van Dale

	What indicates about the LU	**Typical label**	**Typical labelling...**	
Currency	In the dimension of time, its use is..	Obsolete Old-fashioned	Greensward Jolly good	✓
Domain	It is used when the subject of discussion is...	Architecture Music	Transept Arpeggio	✓
Evaluation	It indicates the speaker or writer's attitude to be...	Pejorative Appreciative	Skinny Slender	✓
Figuration	The type of meaning it holds is...	Lit(eral) Fig(urative)	Rich man Rich reward	✓
Region	It is mainly used in...	American British	Sidewalk Pavement	✓
Register	Its use indicates a ..manner of speech/writing	Informal Formal	Shut up! Be silent!	✓
Status	It is non standard language belong to the subset...	Slang Dialect	The nick (prison) A bonny lassie	✓
Style	It is normally used in a...text	Poetic Technical	casement throughput	✓
Specificity	It is used by people in the...world	Military Medical	Anti-personnel Intra-uterine	✓
Usage	[restriction, pragmatics, real-world information etc.]	Offensive	Racist, sexist &c. terms, abuse	✓

3.2 Computational Lexicons

3.2.1 Collins-Robert English-French Lexical-Semantic Database

3.2.1.1 Description

The Collins-Robert Lexical-Semantic Database has been developed by Thierry Fontenelle and his team at the University of Liège (BE) on the basis of the machine-readable version of the Collins-Robert English-French dictionary (1978). The database (described in Fontenelle, 1997), shows the feasibility of the use and exploitation of bilingual lexical resources available in machine-readable form. The source material has been enriched (mostly hand-writing encoding) with lexical-semantic information following Mel'čuk's descriptive apparatus of lexical functions. The source material, enriched, was introduced in a relational database. This database contains around 70,000 pairs of collocates and semantically related items.

For accessing the database, there exists a retrieval program and a command line interface which allows the user to parametrise the information required. These parameters allow the user to query the database by supplying access keys. These access keys refer to the following:

-i: italicized metalinguistic item (appears in italics in the printed dictionary: collocation, typical subject, typical object, synonym, etc)
-h: English headword (in the printed version of the dictionary)
-pos: part of speech of the English headword
-lex: lexical function linking the italicized item and the headword (the mechanism and the list of lexical functions can be found in Chapter 5 of the book referred to above)

The results of these queries are shown in the following format, which will be used in the examples quoted below.

1 (2) : ˜3˜ => 4 <5> (6,7)

(1) English headword
(2) PoS of the English headword
(3) italicized item
(4) French translation of the headword
(5) morphosyntactic features of the French word
(6) French translation of the italicised item
(7) The standard lexical function or lexical-semantic relationship.

The Collins-Robert database source material were the tapes of the printed version of the bilingual English (SL)-French (TL) dictionary, unidirectional. The derived database, enriched with the Lexical Functions characterisation of the different entries, is an active dictionary. The documentation stresses that collocational information in the original dictionary was included specifically to help speakers of the source language (English) select the best target language (French) equivalent of the headword.

As a final comment before going into the description table, it should be noted that this database should be thought as an information resource where there is no real encoding of most of the categories mentioned in the table, but where these can be derived thanks to the Lexical Functions annotation, and the relation between the different entries for a headword, for the presence of a given wordform in the called italicised word, etc.

3.2.1.2 Lexical-Semantic annotation

Most of Mel'čuck's Lexical Functions are used, and new functions have been included. We provide here an example of one lexical funcion.

"Mult" is a function that takes as argument the italicised word and gives the headword:

Mult(state)= confederacy

The italicised word in the original can however correspond either to a collocate or to a related word which however is unlikely to appear together with the headword:

Mult(bee) = cluster

Mult(bee) = swarm

swarm (n) : ~bee~ => essaim <m> (abeille,mult)
cluster (n) : ~bee~ => essaim <m> (abeille,mult)

3.2.1.3 Synoptic table of information types in the Collins-Robert Lexical-Semantic Database

Table 9.: Lexical Information in the Collins-Robert Lexical-Database

	Entry component		Present	Comments
1	headword		✓	
2	phonetic transcription			
3	variant form			Variants are not linked to the headword although there exist different SL headwords. For TL some information is also available.
4	inflected form		✓	Inflected forms of the SL and TL when special translation. See note 1 in 3.2.1.4.
5	Cross-reference			Present in the DB but not accessible
6	**Morphosyntactic Information**			
	a	Part-of-speech marker	✓	For English headword
	b	Inflectional class		
	c	Derivation	Not encoded explicitly	The lexical functions can help to find this information. See note 2.
	d	Gender	✓	Gender is marked in TL wordforms when relevant.
	e	Number	✓	Number is marked in TL wordform with a tag and for the SL the wordform is given as information contained in the headword entries
	f	Mass vs. Count		Not encoded as such although the lexical functions 'mult' and 'sing' for some items refer to this distinction.
	g	Gradation		
7	Subdivision counter			
8	Entry subdivision			Senses of a given headword are listed according to translation requirements when a headword is queried.
9	Sense indicator		✓	Information is supplied by means of the italicised wordforms
10	linguistic label		✓	Some tags are used such as: Informal, liter, euph.
11	**Syntactic Information**			
	a	Subcategorization frame		PoS tags include the reference to the basic valence of the verb, i.e. vt, vi. Information on bound prepositions and on phrasal verbs particles also mentioned for some verbs depending on its relevance for translation.

	b	Obligatority of complements		
	c	Auxiliary		Not encoded specifically although for some words whose translation into French results into an adjective or participial phrase is specified.
	d	Light or support verb construction	✓	Specified by means of italicised text and Lexical Functions. See note 3
	e	Periphrastic constructions		
	f	Phrasal verbs	✓	Particles are included in the verbal entry according to translation needs.
	g	Collocator	✓ By means of Lexical Functions	*(i.) typical subject /object of verb, noun modified by adjective etc.* *(ii.) type of collocation relation represented*
	h	Alternations		It can be inferred from *vt* vs. *vi* specification of the same headword and related lexical functions See note 4
12	**Semantic Information**			
	a	Semantic type		However, some links between words can be traced. See note 5
	b	Argument structure		
	c	Semantic relations	✓	encoded as lexical functions, see note 6
	d	Regular polysemy		
	e	Domain	✓	Some references as relevant for translation
	f	Decomposition	✓	Some lexical functions refer to these meaning components
13	Translation		✓	
14	Gloss		✓	
15	Near-equivalent			
16	Example phrase (straightforward)			
17	Example phrase (problematic)		✓	See note 7.
18	multiword unit		✓	
19	subheadword *also* secondary headword			
20	usage note			
21	Frequency			

3.2.1.4 Notes

1. Inflected forms and morphosyntactic information

ability (n) : ~power~ => aptitude <f> (<to> <do> … faire) (pouvoir,syn)
ability (n) : ~proficiency~ => aptitude <f> (<to> <do> … faire) (aptitude,syn)
ability (n) => capacité <f> (<to> <do> pour faire)
ability (n) => compétence <f> (<in> en, <to> <do> pour faire)
ability (n) : ~cleverness~ => habileté <f> (intelligence,syn)
ability (n) => talent <m>
ability (n) {ABILITIES} : ~mental powers~ => talents <mpl> (capacities mentales,gener)
ability (n) {ABILITIES} => dons intellectuals

2. Derivational morphological information can be derived from the relation between the italicised word and the headform when the Lexical Function is marked as A0, A1, S0, Adv0, Able1, V0.

Examples:

life (n) : ~live~ => vie <f> (vivre,s0)

professional (adj) : ~profession~ => professionnel (profession,a0)
impulsive (adj) : ~impulse~ => impulsif (impulsion,a1)

suspiciously (adv) : ~suspicion~ => avec m,fiance (soupçon,adv1)

practise (vt) : ~practice~ => pratiquer (pratique,v0)

However this tag is not exclusive of the morphological derivation relation, but it includes semantic relations and so we can also find other examples such as the followings:

capacity (n) : ~hold~ => contenance <f> (contenir,s0)
loan (n) : ~borrowed~ => emprunt <m> (emprunt,s0)
vulgar (adj) : ~common people~ => vulgaire (peuple,a0)

free (adj) : ~liberty~ => libre (libert,a1)

free (adj) : ~liberty~ => autonome (liberté,a1)

credit (n) : ~belief~ => ajouter foi … (croyance,v0)

3. Support verbs.

The Mel'cuk lexical function *Oper* roughly correspond to Gross 'support verb' idea. Although verbs mentioned are not only those that have no lexical meaning. Compare the following samples to have an idea of the coverage of this Lexical Function (which can be used in combination with other lexical functions).

Example: *mistake*

make (vt) : ~mistake~ => faire (erreur,oper1)
let past (vt sep) : ~mistake~ => laisser passer (erreur,permoper1)
commit (vt) : ~mistake~ => commettre (erreur,oper1)
overlook (vt) : ~mistake~ => laisser passer (erreur,permoper1)

Example: *attention*

engross (vt) : ~attention~ => absorber (attention,oper2)
excite (vt) : ~attention~ => exciter (attention,oper2)
fix (vt) : ~attention~ => fixer (attention,oper1)
divert (vt) : ~attention~ => détourner (attention,finoper1)
draw (vt) : ~attention~ => attirer (attention,oper2)
engage (vt) : ~attention~ => éveiller (attention,incepoper2)
invite (vt) : ~attention~ => demander (attention,oper2)
occupy (vt) : ~attention~ => occuper (attention,oper2)
focus (vt) : ~attention~ => concentrer (<on> sur) (attention,oper1)
arrest (vt) : ~attention~ => retenir (attention,oper2)
capture (vt) : ~attention~ => capter (attention,oper2)
turn (vt) : ~attention~ => tourner (attention,oper1+oper2)
win (vt) : ~attention~ => capter (attention,oper2)
crave (vi) : ~attention~ => solliciter (attention,oper2)
claim (vt) : ~attention~ => demander (attention,oper2)
concentrate (vt) : ~attention~ => concentrer (<on> sur) (attention,oper1)
take up (vt sep) : ~attention~ => occuper (attention,oper2)
switch (vt) : ~attention~ => reporter (<from> de, <to> sur) (attention,incepoper1)

4. Alternations

Causative/inchoative alternation can be extracted from the information of the PoS tag and the lexical function which includes the 'caus' function in the vt case and not in the vi case. This treatment allows a possible correlation of this property with the verb's belonging to one of the many sub-classes of change of state verbs: verbs of sound/noise, cooking, verbs of impairment, etc.

ring (vt) : ~bell~ => (faire) sonner (cloche,causson)
ring (vi) : ~bell~ => sonner (cloche,son)
toll (vt) : ~bell~ => sonner (cloche,causson)
toll (vi) : ~bell~ => sonner (cloche,son)
sound (vt) : ~bell~ => sonner (cloche,causson)
sound (vi) : ~bell~ => sonner (cloche,son)
chime (vt) : ~bell~ => sonner (cloche,causson)
chime (vi) : ~bell~ => carillonner (cloche,son)
peal (vt) : ~bell~ => sonner (… toute volée) (cloche,causson)
peal (vi) : ~bell~ => carillonner (cloche,son)

5. Semantic type.

Depending on the translation requirements, some headforms are distinguished by the italicised wordform and the lexical function in such a way that semantic typing could be derived.

leg (n) : ~horse~ => jambe <f> (cheval,part)
leg (n) : ~person~ => jambe <f> (personne,part)
leg (n) : ~bird~ => patte <f> (oiseau,part)
leg (n) : ~insect~ => patte <f> (insecte,part)
leg (n) : ~animal~ => patte <f> (animal,part)
leg (n) : ~lamb~ => gigot <m> (agneau,part)
leg (n) : ~beef~ => gŒte <m> (boeuf,part)
leg (n) : ~veal~ => sous-noix <f> (veau,part)
leg (n) : ~chicken~ => cuisse <f> (poulet,part)
leg (n) : ~frog~ => cuisse <f> (grenouille,part)
leg (n) : ~pork~ => cuisse <f> (porc,part)
leg (n) : ~venison~ => cuissot <m> (venaison,part)
leg (n) : ~table~ => pied <m> (table,part)
leg (n) : ~stocking~ => jambe <f> (bas,part)
leg (n) : ~trousers~ => jambe <f> (pantalon,part)
leg (n) : ~journey~ => étape <f> (voyage,part)

6. Lexical-Semantic relations:

The Lexical Functions which allow the relation of different wordforms are:

hyperonymy : gener

synonymy: syn

antonymy: anti

Example: *farm*

holding (n) : ~farm~ => propriété <f> (ferme,syn)
home (cpd) [HOMESTEAD] : ~farm~ => ferme <f> (ferme,syn)
grange (n) : ~farm~ => ferme <f> (ferme,syn)

7. Examples of problematic cases

lay (vt) {TO LAY BARE ONE'S INNERMOST THOUGHTS} => mettre á nu ses pensées les plus profondes
lay (vt) {TO LAY BARE ONE'S INNERMOST THOUGHTS} => dévoiler ses pensées les plus profondes
lay (vt) {TO LAY BARE ONE'S INNERMOST FEELINGS} => mettre á nu ses sentiments les plus secrets
lay (vt) {TO LAY BARE ONE'S INNERMOST FEELINGS} => dévoiler ses sentiments les plus secrets
lay down (vt sep) {TO LAY DOWN ONE'S ARMS} : ~give up~ => déposer ses armes (abandonner,syn)
lay down (vt sep) {TO LAY DOWN ONE'S ARMS} => déposer les armes
lay off (vt fus) {LAY OFF (IT)!} : ~stop~ => tu veux t' arrèter? [informal] (arrèter,imper)
lay off (vt fus) {LAY OFF (IT)!} : ~touch~ => touche pas! [informal] (toucher,antiimper)
lay off (vt fus) {LAY OFF (IT)!} => pas touche! [very informal]
lay off (vt fus) {LAY OFF (IT)!} => bas les pattes! [very informal]

3.2.2 The FrameNet Lexicon Database

This is a summary of the information types in the FrameNet database (Baker et al., 1998), which may be queried on http://163.136.182.112/framesql/notes/index.html. Further information is to be found on the FrameNet home page: http://www.icsi.berkeley.edu/~framenet/index.html.

The only available FrameNet data at the moment is for English, but parallel work in German is currently in progress for some of the frames; FrameNet is planning to start a similar analysis of Japanese, while an associated initiative in Hong Kong proposes to analyse Mandarin and Cantonese: these parallel monolingual databases will feed into an MT system, the semantic annotations essentially forming an interlingua. This is explained in Note **J**.

We summarize the information using the grid "Lexical information in bilingual resources" , and in the summary table we have retained the first two columns. '#' in column 1 means that this is an addition to the table.

FrameNet makes no attempt to record all the standard dictionary information relating to a word form (e.g. phonetic transcription, variant forms, morphosyntactic information, etc.). Entries in the FrameNet lexicon record the linked semantic and syntactic valences of a lemma (a word in one of its senses) by means of manually inserted annotation tags. These link the frame-based semantic roles (or 'frame elements', FEs) to their syntactic expression in the immediate grammatical context of the word occurring in a corpus sentence. The syntactic information recorded is twofold: (1) the phrase type (PT) of the word or words being annotated, and (2) the grammatical function (GF) of that word or phrase within the context of the lemma. The annotation of these sentences is therefore tripartite, e.g. (from the British National Corpus) *Princess Diana and Prince Charles have admitted in writing their marriage is in trouble* is annotated thus (FrameNet annotations in bold, others from the BNC):

<S TPOS="31880343"> **<T TYPE="sense1"> </T> <C FE="Spkr" PT="NP" GF="Ext">***Princess*/NP0 *Diana*/NP0 *and*/CJC *Prince*/NP0 *Charles*/NP0 **</C>** *have*/VHB **<C TARGET="y">** *admitted*/VVN **</C> <C FE="Medium" PT="PPing" GF="Comp">***in*/PRP *writing*/VVG **</C> <C FE="Msg" PT="Sfin" GF="Comp">***their*/DPS *marriage*/NN1 *is*/VBZ *in*/PRP *trouble*/NN1 **</C>** ./PUN </S>

The names of the frame elements in this sentence, SPEAKER, MEDIUM and MESSAGE, are transparent: they belong to the COMMUNICATION/STATEMENT frame and their full database form is COMMUNICATION/STATEMENT/SPEAKER etc. A fuller description of a FrameNet lexical entry is given in note **B** below, where the valence patterns are shown in detail.

3.2.2.1 Synoptic table of information types in the FrameNet lexicon.

Table 10: Lexical Information in the FrameNet lexicon

	Entry component		Present	FrameNet data	See note ...
1	headword		✓	Lexical unit (a lemma in one of its senses)	
2	Phonetic transcription		✓		
3	variant form		✓	variant forms are linked in the database	
4	inflected form				
5	Cross-reference				
6	**Morphosyntactic information**				
	a	Part-of-speech marker	✓	p-o-s is a component in the lexical entry	
	b	Inflectional class			
	c	Derivation			
	d	Gender			
	e	Number			
	f	Mass vs. Count			
	g	Gradation			
7	Subdivision counter		✓	lemma (see 1) = wordform + sense number	
8	Entry subdivision			(a FrameNet entry cannot be subdivided)	
9	Sense indicator		✓	The most relevant definition from Concise Oxford Dictionary is included in each lexical entry in order to help the human user identify the sense of the lemma.	
10	linguistic label				
11	**Syntactic information**				
	a	Subcategorization frame	✓	exhaustively covered and linked to semantic roles	B
	b	Obligatority of complements	✓	No attempt is made to declare the intuition that something is obligatory: FrameNet simply tries to record the tings that occur	B
	c	Auxiliary			
	d	Light or support verb construction	✓	support verb: indicates relationships similar to Mel'cukian functions	C
	e	Periphrastic constructions			
	f	Phrasal verbs	✓	Minimally, and not systematically, covered in FrameNet; the verb+particle unit is linked with a tag target-mate, but current software does not extraxt these as separate entries	

	g	Collocator		this concept belongs to a bilingual dictionary targeting human users: it should not be confused with 'collocation' as discussed in corpus literature; it has no place in FrameNet	
	h	Alternations	✓	exhaustively covered and linked to semantic content	B
12	**Semantic information**				
	a	Semantic type		not in FrameNet: links to WordNet synsets were planned but proved impossible to implement	
	b	Argument structure	✓	exhaustively covered and linked to syntactic expression	B
	c	Semantic relations	✓	each lemma is linked to its immediate semantic neighbours by belonging to the same frame	A, D
	d	Regular polysemy		not covered currently, but the FrameNet database is an ideal environment for identifying instances of regular polysemy	
	e	Domain		FrameNet's 'domain' is not the regular subject-field type normally found in dictionaries and lexicons, see note.	A
	f	Decomposition			
#	**Syntactico-semantic information**				
	#a	N(P)+N compounds		semantic relationship between noun and its modifier is shown in terms of the FEs involved	E
	#b	non-instantiated semantic roles		frame elements which are understood in the sentence but not overtly expressed	F
	#c	frame-wide lexical instantiations of semantic roles		see note	G
	#d	semantic roles of prepositions		see note	H
##	**Lexical semantic information**				
	##a	corpus profiles of lexical items		see note	I
13	Translation		✓	this can be derived	J
14	Gloss				
15	Near-equivalent				
16	Example phrase (straightforward)		✓	each valence pattern comes complete with sentences extracted from British National Corpus, including the location of the keyword in the BNC, and annotated with respect to the top-level phrases accompanying a given target..	B
17	Example phrase (problematic)			this concept belongs to a bilingual dictionary targeting human users	

18	multiword unit		not in current FrameNet but will be included in Phase 2	
19	subheadword *also* secondary headword			
20	usage note			
21	Frequency	✓	not yet available but automatic assignment of Frame Elements to lemmas in raw corpus data is a current objective and when successful will allow computation of absolute and relative frequencies of lemmas (word senses) in the corpus.	

3.2.2.2 Notes

A. [12c, 12e] Semantic Relations, and Domains

The FrameNet lexicon is currently subdivided into the following semantic domains: BODY, COGNITION, COMMUNICATION, EMOTION, GENERAL, HEALTH, LIFE, MOTION, PERCEPTION, SOCIAL, SPACE, TIME, TRANSACTION.

Each domain is further subdivided into various frames, for instance, the COMMUNICATION domain includes : CANDIDNESS, COMMITMENT, CONVERSATION, ENCODING, GESTURE, HEAR, MANNER, NOISE, QUESTIONING, REQUEST, RESPONSE, STATEMENT, VOLUBILITY.

A semantic frame is a script-like structure of inferences, linked by linguistic convention to the meanings of linguistic units (lemmas). Each frame identifies a set of frame elements (FEs) - participants and props in the frame. A frame semantic description of a lexical item identifies the frames which underlie a given meaning and specifies the ways in which FEs, and constellations of FEs, are realized in structures headed by the word.

Domain and frame names are all provisional. Once the initial lexical entries have been compiled for the basic vocabulary of the language, there will be a process of harmonization in which many labels may be changed.

More domains and frames will be added during Phase 2 of the project.

B. [11a, 11b, 12h] A lexical entry: semantic & syntactic valence links

Here is a summary of the lexical entry for the verb *drawl*, within the Communication/Manner frame. Other lemmas in this frame are: *babble, bluster, chant, chatter, gabble, gibber, jabber, lisp* etc. The entry starts by listing the FE set (= subset of the frame elements used in the annotation and description of drawl): ADDRESSEE, DEPICTIVE-ACTOR, MANNER, MESSAGE, SPEAKER.

The main part of the entry consists of a set of linked semantic and syntactic valences (shown in the table below). The column heading '**frequency'** is something of a misnomer, as the figures in that column simply indicate the number of annotated sentences for each pattern in the FrameNet lexicon, and bear no systematic relationship to frequency in the corpus. FrameNet aims simply to include all the patterns found in the BNC. The term **pattern** refers to a configuration of frame elements forming a grammatical unit (phrase,

clause, or sentence): the various syntactic ways in which the pattern is realised are listed below each. In the database the annotated sentences are listed separately via hypertext links from the frequency numbers; however for ease of reference I have included in italics one example sentence from the corpus for each syntactico-semantic pattern:

freq (40)	**Patterns**
20	**Message** **Speaker**
02	**QUO.Comp+ QUO.Comp** **NP.Ext**
	*" Well, well," (*MSG*) drawled Havvie Blaine (*SPKR*), at last , "what have we here?" (*MSG*)*
18	QUO.Comp NP.Ext
	*"Well, well, well," (*MSG*) he (*SPKR*) drawled.*
02	**Message** **Speaker** **Depictive-Actor**
02	**QUO.Comp** **NP.Ext** **PP_with .Comp**
	"If you say so," (MSG) *he (*SPKR*) drawled, with a smug expression (*DEP-ACT*).*
03	**Message** **Speaker** **Manner**
01	**NP.Ext** **CNI** **AVP.Comp**
	*It was a cool challenge (*MSG*), drawled* CNI *(*SPKR*) so quietly (*MANR*) that she almost missed it.*
02	QUO.Comp NP.Ext AVP.Comp
	*"You're very liberal with your criticism," (*MSG*) he (*SPKR*) drawled huskily (*MANR*).*
05	**Speaker**
05	**NP.Ext**
	*Luke (*SPKR*) drawled, allowing a weary sigh to escape from his lips.*

01	**Speaker** **Addressee**
01	**NP.Ext** **PP_to .Comp**
	*Fonda studiously ignores the hairs as he (*SPKR*) drawls to an off-screen interrogator (*ADD*).*
03	**Speaker** **Manner**
03	NP.Ext AVP.Comp
	*She (*SPKR*) talked a little to herself, lowering her voice and drawling carefully (*MANR*).*
01	**Speaker** **Manner** **Message**
01	**NP.Ext** **PP_in .Comp** **QUO.Comp**
	*His mouth twisted slightly as he (*SPKR*) drawled in a sardonic tone (*MANR*), "What's the matter?" (*MSG*)*
04	**Speaker** **Message**
03	**NP.Ext** **NP.Obj**
	*He (*SPKR*) drawled the warning (*MSG*).*
01	NP.Ext Sfin.Comp
	*Linley (*SPKR*) drawled that there was nothing to get upset about (*MSG*).*
01	**Speaker** **Message** **Manner**
01	**NP.Ext** **NP.Obj** **PP_with .Comp**
	*Jackson (*SPKR*) drawled the word (*MSG*) with a slow complacency (*MANR*).*

C [11d] Support verbs

FrameNet defines 'support verb' in very broad terms, and the links recorded are close to Mel'cukian functions. Here are some examples of the rich collocational information thus recorded:

1. For the nouns *allegation* and *announcement* in the BNC, a query relating to verbs annotated as support verb shows only *make,* e.g.
He said he would make an announcement about his plans.
A teacher was summarily dismissed after making allegations against her colleagues.

2. For the noun *complaint* the following support verbs are recorded (underlined type in the examples below) with the same function as *make*:
Members of third parties may make complaints in writing.
I wondered if he'd registered a complaint against you.
A north-east woman has lodged a complaint after an ambulance took almost an hour to arrive at an accident.
There are a few who express complaints, with the quality of care offered.
In these discussions the boys often voice similar complaints to the girls.
After his release he submitted a formal written complaint to the Procurator General's Office.
I have no complaints with your work.
In spite of complaints brought by leaders of trade unions ...

D [12c] Semantic relations

Thus the lemmas so far recorded as belonging to the COMMUNICATION / CATEGORIZATION frame include: *categorization* n, *categorize* v, *characterization* n, *characterize* v, *class* v, *classification* n, *classify* v, *construe* v, *define* v, *definitin* n, *depict* v, *depiction* n, *describe* v, *description* n, *interpret* v, *interpretation* n, *perceive* v, *portray* v, *redefine* v, *redefinition* n, *regard* v, *represent* v, *representation* n, ... etc. etc.
These lemmas are linked by the fact that the same set of frame elements is used in recording their valence patterns.

E [#a] N+N compounds

For the noun *allegation* in the COMMUNICATION / STATEMENT frame, two types of relationship are recorded:

1. where the modifying N is annotated as MESSAGE, as in:
child abuse allegations
assault allegations
corruption allegations
ballot-rigging allegations
torture allegations
forgery allegatioon
conspiracy allegation
espionage allegations

2. where the modifying N or NP is annotated as SPEAKER, as in:
government allegations that ...
newspaper allegations of ...
the Thatcher allegations about ...

F [#b] Non-instantiated semantic roles ('frame elements')

FrameNet records three distinct types of semantic elements which are not lexically realized in the sentence:

1. CNIs (constructional null instantiations)

i.e. licensed by the grammar of the language, as in:

There are briefs and de-briefs, and their efforts in the skies are closely scrutinised and criticised.

In this case the annotated sentence is as follows, showing that the frame element JUDGE is not expressed, since the grammar of the language allows passives without expression of the subject of the active verb:

<S TPOS="101969240"> *There*/EX0 *are*/VBB *briefs*/NN2 *and*/CJC *de-briefs*/NN2 ,/**PUN** **<C FE="Eval" PT="NP" GF="Ext">**their/DPS efforts/NN2 in/PRP the/AT0 skies/NN2 **</C>** are/VBB closely/AV0 scrutinised/VVN and/CJC **<C TARGET="y">** criticised/VVN **</C>**
<C FE="Judge" PT="CNI"> </C> **<C FE="Reas" PT="INI"> </C>** /PUN </S>

3. INIs (indefinite null instantiations)

i.e. those where a no definite entity has to be known to the interpreter of the sentence if it is to be fully understood, as in:

In particular, the ACE scheme was heavily and repeatedly criticised. Here the frame element REASON is not expressed yet the sentence is understood.

In this case the annotated sentence is as follows:

<S TPOS="48756144"> *In*/AV0 *particular*/AV0 ,/PUN **<C FE="Eval" PT="NP" GF="Ext">** the/AT0 ACE/AJ0 *scheme*/NN1 **</C>** *was*/VBD **<C FE="Manr" PT="NP" GF="Ext">***heavily*/AV0 *and*/CJC *repeatedly*/AV0 **</C>** **<C TARGET="y">** *criticised*/VVD-VVN **</C>** **<C FE="Judge" PT="CNI"></C>** **<C FE="Reas" PT="INI"></C>** ./PUN </S>

4. DNIs (definite null instantiations)

i.e. those where a definite entity (usually expressed in the previous context) has to be known to the interpreter of the sentence if it is to be fully understood, as in:

Who can they blame now?

Here the frame element REASON, although it must be known to both SPEAKER and ADDRESSEE if the message is to be conveyed, is not overtly expressed.

In this case the annotated sentence is as follows:

<S TPOS="106031246"> <T TYPE="sense1"> </T> **<C FE="Eval" PT="NP" GF="Ext">***Who*/PNQ **</C>** *can*/VM0 **<C FE="Judge" PT="NP"GF="Ext">** *they*/PNP **</C>** **<C TARGET="y">***blame*/VVI **</C>** **<C FE="Reas" PT="DNI"> </C>** *now*/AV0 ?/PUN "/PUQ </S>

G [#c] Frame-wide lexical instantiations of semantic roles

A query to the database will produce a listing of lexico-syntactic realisations of semantic roles (frame elements) across the frame. The table below shows how the FE TOPIC is syntactically realized in the COMMUNICATION/QUESTIONING Frame in the context of individual lemmas.

Frame = COMMUNICATION/QUESTIONING

Freq	patterns realizing TOPIC	in the context of these lemmas
99	PP_about .Comp	*grill* n, *inquire* v, *inquiry* n, *interrogate* v, *interrogation* n, *query* n, *question* v, *questioning* n, *quiz* v
23	DNI	*inquiry* n, *query* n
12	PP_on .Comp	*grill* v, *query* n, *question* v, *quiz* n
08	PP_into .Comp	*inquire* v, *inquiry* n
08	PPing_about .Comp	*inquiry* n, *question* n
04	PP_after .Comp	*inquire* v
04	PP_as .Comp	*inquiry* n, *query* n
03	PP_regarding .Comp	*inquiry* n, *query* n
02	PP_in .Comp	*inquiry* n, *question* n
02	PP_over .Comp	*inquiry* n, *quiz* v
01	AJP.Mod	*query* n
01	N.Mod	*query* n
01	NP.Comp+PP_on .Comp	*query* n
01	PP_concerning .Comp	*question* v
01	PP_of .Comp	*inquire* v
01	PPing_on .Comp	*question* n

H [#d] Semantic roles of prepositions

A query to the database will produce listings of the semantic roles of prepositions as heads of PPs, as shown in the table below, which summarizes the behaviour of prepositions in the MOTION/ARRIVING frame :

Frame = MOTION/ARRIVING

Freq	FE	expressed in these patterns	in the context of these lemmas
88	**GOAL**		
25		PP_to .Comp	*approach, arrive, come, entrance, return, visit*

16		PP_into .Comp	*approach, come, enter, entrance, return*
11		PP_at .Comp	*arrive, visit*
08		PP_in .Comp	*arrive, come*
07		PP_back .Comp	*arrive, come*
04		AVP.Comp+PP_to .Comp	*come, return*
04		PP_with .Comp	*visit*
03		PP_over .Comp	*come*
03		PP_round .Comp	*come*
03		PP_up .Comp	*come*
02		PP_down .Comp	*come, visit*
02		PP_on .Comp	*arrive, visit*

76	**SOURCE**		
64		PP_from .Comp	*approach, arrive, come, enter, entrance, return, visit*
07		PP_out .Comp	*come*
04		PP_away .Comp	*come*
01		PP_out .Comp+ PP_from .Comp	*come*
35	**PATH**		
11		PP_via .Comp	*approach, arrive, come, enter*
07		PP_through .Comp	*approach, arrive, enter*
05		PP_by .Comp	*enter*
04		PP_at .Comp	*come, enter*
04		PP_towards .Comp	*come, return*
02		PP_on .Comp	*approach*
01		PP_across .Comp	*return*
01		PP_round .Comp	*come*
35	**THEME**		
21		PP_of .Comp	*approach, entrance, return, visit*
08		PP_by .Comp	*approach, visit*
06		PP_from .Comp	*visit*
12	**VEHICLE**		
10		PP_by .Comp	*arrive, come, return*
01		PP_in .Comp	*arrive*

01		PP_on .Comp	*visit*
11	**COTHEME**		
09		PP_with .Comp	*come, enter, return, visit*
02		PP_along .Comp	*come*
06	**MANNER**		
03		PP_with .Comp	*approach, enter*
02		PP_like .Comp	*come*
01		PP_on .Comp	*approach*

I [##a] Corpus profiles of lexical items

It is possible to derive from the database information about the semantic roles associated with any specific lexical item in the corpus. This is shown below for a small section of the results of a query about *road*, showing that the word occurs as head N of an NP realizing a specific Frame Element in the MOTION domain as follows:

In the MOTION domain the lemma *road* occurs ...

freq	realizing this FE	in this frame	in the context of this lemma	in these patterns
	AREA			
03		TRANSPORTATION	*cruise*	PP_on .Comp
02		SELF-MOTION	*slither*	PP_on .Comp
01		PATH-SHAPE	*swerve*	PP_over .Comp
01		SELF-MOTION	*prowl*	NP.Obj
	GOAL			
01		SELF-MOTION	*waltz*	PP_across .Comp
04		SELF-MOTION	*dash, rush step*	PP_into .Comp
01		PLACING	*inject*	PP_into .Comp
01		PATH-SHAPE	*swing*	PP_into .Comp
01		PATH-SHAPE	*leave*	PP_alongside .Comp
01		PLACING	*install*	PP_at .Comp
01		PLACING	*park*	PP_down .Comp
01		PLACING	*park*	PP_in .Comp
01		SELF-MOTION	*jump*	PP_in .Comp

			+PP_in .Comp
01	SELF-MOTION	*run*	NP.Obj

etc. etc.

J [13] Translation

The equivalent in another language ('translation') would be derived by

2. selecting the appropriate lemma by matching frame element patterns of source and target languages (stored in the FrameNet databases for the various languages) from a candidate list provided by a bilingual or multilingual glossary extracted from a machine-readable bilingual or multilingual dictionary;

3. using the PT and GF syntactic annotations (also stored in the FrameNet databases) to generate grammatical sentences in the target language. This operation (2) is a dynamic process performed on text to be translated, and cannot be stored as part of a static lexicon.

3.2.3 Multilingual information in EuroWordNet and ItalwordNet

In order to define which information is present in the EWN and IWN databases, we will give a brief description of the data structure. Following the synoptic table we will try to determine whether the information commonly found in bilingual dictionaries is present in the data structure and in which form.

3.2.3.1 Description

Within the European project EuroWordNet (Vossen, 1998, Alonge et al., 1998), semantic information was encoded in each of the languages dealt with, in form of lexical semantic relations between synonym sets (the *synsets*, the core of the whole structure, following the WordNet model, Miller et al., 1990).

A rich framework of relations was designed and they have been introduced for their supposed relevance and usefulness in linguistic applications, e. g. Cross part of speech relations.

Synonymy, hyp(er)onymy and xpos relations have been extensively encoded, while the more "sophisticated" relations have been encoded just for selected classes of words

ItalWordNet (IWN), the Italian follow-up of EWN, is a part of a National project (SI-TAL, Integrated System for Automatic Treatment of Language) which aims at building various integrated language resources for the automatic treatment of the Italian written and spoken language.

In ItalWordNet we are now extending the WordNet produced for Italian during the previous project, extensively inserting adjectives, adverbs, multiword expressions and instances, and increasing the number of present nouns and verbs (with the goal of 50,.000 total lemmas) (Roventini et al., 2000).

A few semantic relations have been added to the previous set, mainly to be used to encode data on adjectives and the EWN Top Ontology has been revised to better represent this part of speech (Alonge et al., 2000) (in EWN, adjectives and adverbs were already present, but just as target of relations from nouns and verbs).

One of the most relevant aspect of E(/I)WN is its multilinguality: each wordnet is linked with all the other language specific wordnets by means of an interlingual index (ILI).

Due to its importance in computational applications, a domain specific wordnet is also being built.

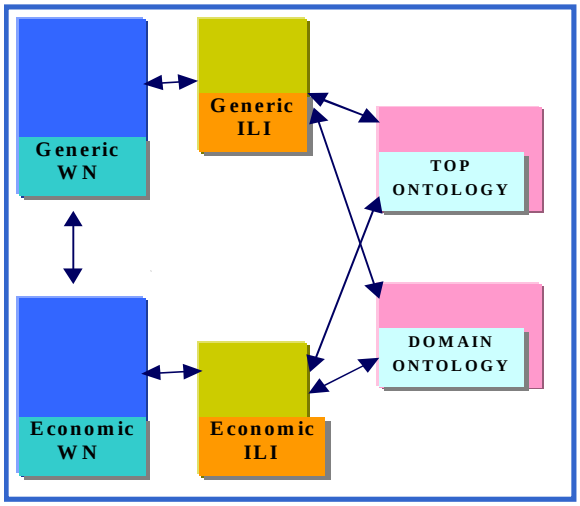


Fig. 5: The overall architecture of the IWN database

The IWN database (see fig. 5) is constituted by:

1. a generic monolingual wordnet;

2. a (generic) Interlingual Index (ILI) (an unstructured version of the Princeton WordNet –1.5- containing all the synsets belonging to this version but not the relations among them). All the synsets of the monolingual wordnet are linked to this "interlingua", to make the resource usable in multilingual applications;

   e.g.:

Dog	Noun	"a member of the genus canis"	1422174
Cad, bounder, blackguard, dog, hound, heel	Noun	"someone who is morable reprehensible"	5980708
Pawl, detent, click, dog	Noun	"a hinged device that fits into a notch of a ratchet.."	5861550

   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.

3. a terminological wordnet, containing synsets found in the economical-financial domain;

4. a terminological ILI, containing synsets partly extracted from WN1.6;

5. the Top Ontology (TO), the hierarchy of language independent concepts reflecting fundamental semantic distinctions;

6. the Domain Ontology (DO), containing a set of domain labels. In EWN this module was only partially developed and used to encode information on computer terminology, whereas in IWN a complete set of labels is being developed.

The following picture (fig. 6) shows an example of the monolingual net surrounding the synset {cane 1} (dog) and its links with the ILI. Dog is also linked, by means of the corresponding base concept, to the Top Concepts of the Ontology.

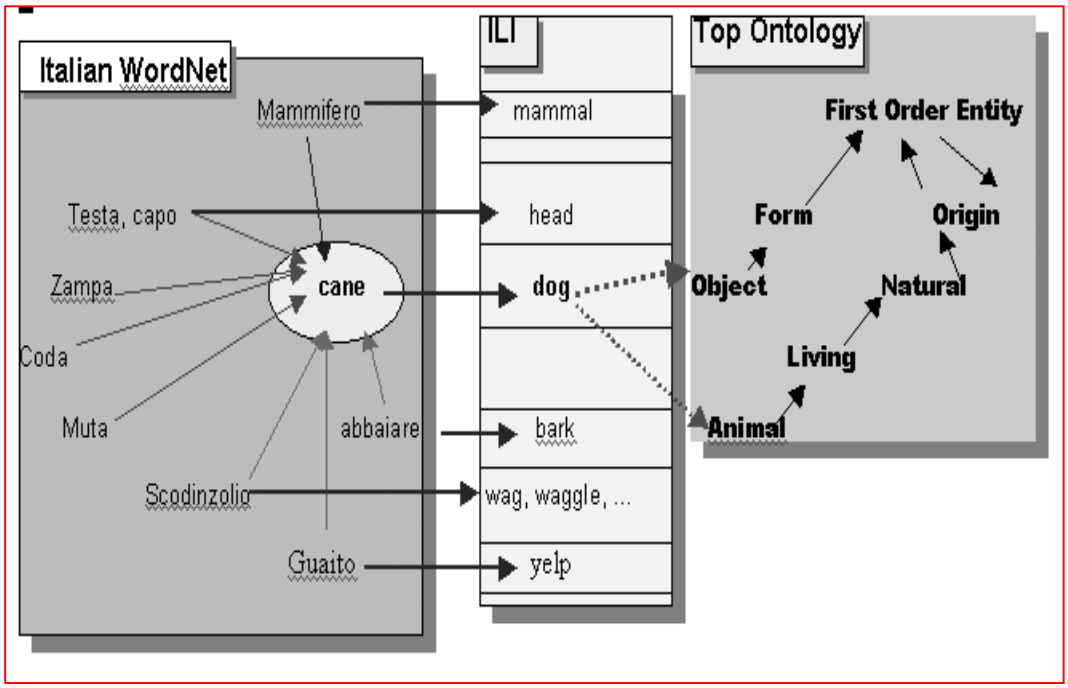


Fig. 6: a synset with its mono/multilingual links

3.2.3.2 Language dependent/language independent information

The information encoded in EWN and IWN can be classified in two ways:

language dependent:

synset level

- synonyms belonging to the synset
- synset POS

- monolingual semantic relations to other synsets of the net (+ information like meaning disjunction, features negation, reversibility)
- interlingual semantic relations to the ILI

variants level (a variant is a member of a synset)

- sense number
- style, usage and domain information
- feature (case, collective, connotation, countability, determiner, infinite clause, finite clause, gender, nominal complement, number, person, tense)
- semantic relation between variants and not between synsets (derivation)

Language independent

Information present in the following modules:

- ILI

  -Relations between the ILI and the Top Ontology

  -Relations between the ILI and the Domain Ontology

- Top Ontology

  -Relations between Top Concepts

- Domain Ontology

  -Relations between Domain Concepts

3.2.3.3 Monolingual/multilingual information

The multilingual link is realized, as we already saw, by means of the interlingual structure that allows the passage from the Italian net to all the other monolingual, language specific wordnets built during EWN.

A semantic relation of equivalence, representing a sort of "relation of translation", links each synset of the net to the ILI. We give the list of all the possible situations and the adopted solutions:

1. the meaning of the Italian synset exactly corresponds to the meaning of an ILI synset

   Between the Italian and the ILI synset an equivalent synonymy relation is established (eq_synonym)
   e.g.:
   Mammifero 1 **eq_synonym→***mammal* 1

2. the meaning is present in the ILI but it doesn't exactly match because:

   a. it was differently classified in WordNet and it has a different definition
      e.g.:
      Dissodamento 1 (l'operazione del dissodare la terra) **eq_near_synonym** *tillage*
      (the cultivation of soil for raising crops)

   b. there is no a *one to one* relationship between the ILI and the Italian sense,
      e.g.:
      coperchio 1 (translation: *lid, cap, cover, top*) **eq_near_synonym** lid
      eq_near_synonym *cap*
      eq_near_synonym *cover, top*

   In these cases, among the Italian and the ILI synsets more than one equivalent near synonymy relations are established (eq_near_synonym)

3. the meaning doesn't exist in the American-English of the ILI (it is a genuine linguistic gap)
   e.g.:
   {*abbacchiare* 1, *bacchiare* 1} (vigorously hitting the branch of a tree with a cane called "bacchio" to make the fruits fall down) **eq_has_hyperonym** *hit*

4. the meaning was not inserted in the WordNet1.5 database.

   {*saldatura* 1, *saldamento* 1} (translation: *welding*) **eq_has_hyperonym** *operation*
   eq_is_caused_by *to weld*

   In the cases 3 and 4, an equivalent hyperonymy relation is codified; in ItalWordNet there are 11 equivalence relations and it's given the possibility to encode complex relations when the eq_(near_)synonymy is not avaible (for example equivalent meronymy, equivalent role, equivalent causes relations and so on..)

3.2.3.4 Examples of IWN entries

The following are some examples of ItalWordNet entries belonging to different part of speech. Each entry is followed by an example of the way it is displayed in the new IWN navigation tool developed at IRST.

3.2.3.4.1 Nouns

1. informatica (computer science)

```
WORD_MEANING ID="n@10393@" PART_OF_SPEECH="n">
 <VARIANTS>
  <LITERAL LEMMA="informatica" SENSE="1" STATUS="new"> </LITERAL>
 </VARIANTS>
 <INTERNAL_LINKS>
  <RELATION R_TYPE="has_hyperonym" ID="IR281">
   <TARGET_CONCEPT ID="n@11231@" PART_OF_SPEECH="n">
    <LITERAL LEMMA="scienza" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
 </INTERNAL_LINKS>
 <EQ_LINKS>
  <EQ_RELATION R_TYPE="eq_synonym" ID="ER282">
   <TARGET_ILI ID="ILI283" PART_OF_SPEECH="n" WORDNET_OFFSET="04084575">
</TARGET_ILI>
  </EQ_RELATION>
 </EQ_LINKS>
</WORD_MEANING>
```

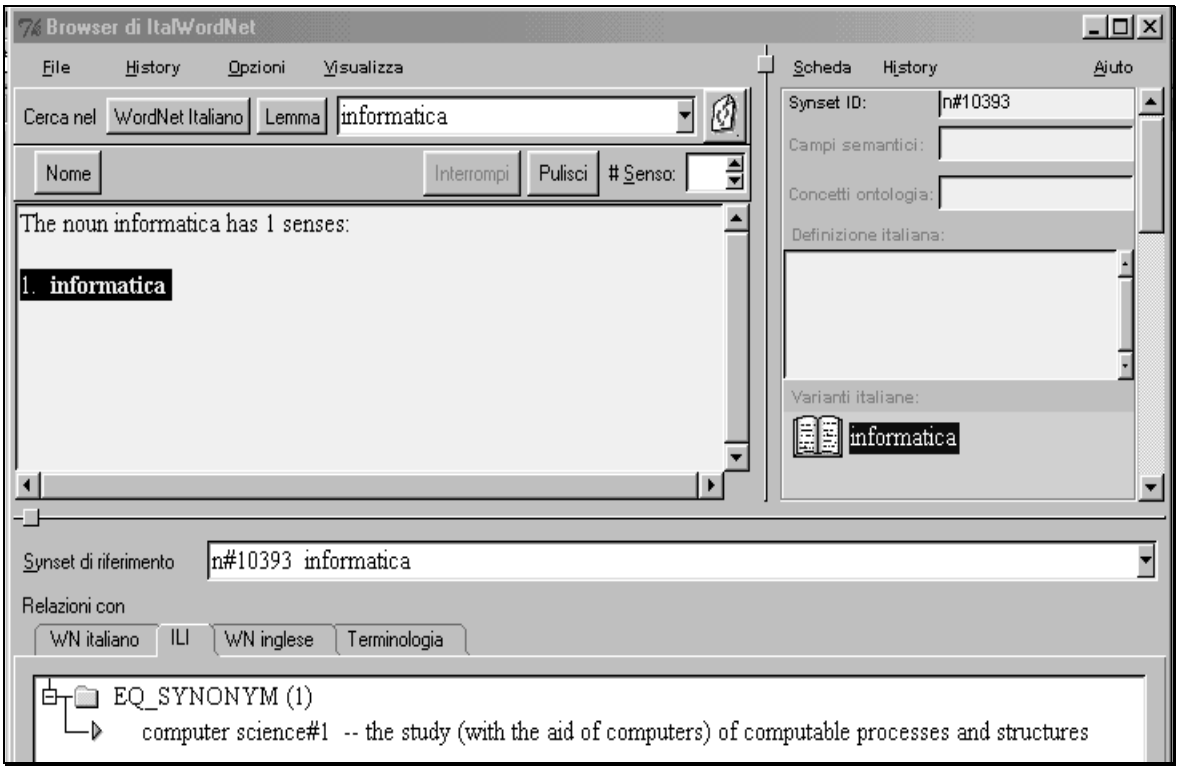


Fig. 7: Example of an IWN entry: "informatica"

2. insufficienza, carenza etc.. (lack)

```
<WORD_MEANING ID="n@10395@" PART_OF_SPEECH="n">
 <VARIANTS>
  <LITERAL LEMMA="insufficienza" SENSE="1" STATUS="new"> </LITERAL>
  <LITERAL LEMMA="carenza" SENSE="1" DEFINITION="mancanza insufficienza." STATUS="Fra-
corpus"> </LITERAL>
  <LITERAL LEMMA="mancanza" SENSE="1" DEFINITION="il mancare." STATUS="new">
</LITERAL>
  <LITERAL LEMMA="deficienza" SENSE="1" STATUS="new"> </LITERAL>
  <LITERAL LEMMA="penuria" SENSE="1" DEFINITION="insufficienza di cose o di persone necessarie."
STATUS="new"> </LITERAL>
  <LITERAL LEMMA="scarsità" SENSE="1" STATUS="new"> </LITERAL>
  <LITERAL LEMMA="assenza" SENSE="2" DEFINITION="il mancare" STATUS="Fra-corpus"
EXAMPLES="'vita in assenza di ossigeno'"> </LITERAL>
  <LITERAL LEMMA="difetto" SENSE="3" EXAMPLES="'difettare di qualcosa'"> </LITERAL>
  <LITERAL LEMMA="strettezza" SENSE="3"> </LITERAL>
  <LITERAL LEMMA="modestia" SENSE="3"> </LITERAL>
  <LITERAL LEMMA="pochezza" SENSE="1"> </LITERAL>
  <LITERAL LEMMA="ristrettezza" SENSE="3"> </LITERAL>
 </VARIANTS>
 <INTERNAL_LINKS>
  <RELATION R_TYPE="xpos_near_synonym" ID="IR287">
   <TARGET_CONCEPT ID="v@4973@" PART_OF_SPEECH="v">
    <LITERAL LEMMA="mancare" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  <RELATION R_TYPE="xpos_fuzzynym" ID="IR288">
   <TARGET_CONCEPT ID="a@42180@" PART_OF_SPEECH="a">
    <LITERAL LEMMA="modesto" SENSE="2"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  <RELATION R_TYPE="has_hyperonym" ID="IR289">
   <TARGET_CONCEPT ID="n@27127@" PART_OF_SPEECH="n">
    <LITERAL LEMMA="stato" SENSE="2"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
 </INTERNAL_LINKS>
 <EQ_LINKS>
  <EQ_RELATION R_TYPE="eq_synonym" ID="ER290">
   <TARGET_ILI ID="ILI291" PART_OF_SPEECH="n" WORDNET_OFFSET="08731035">
</TARGET_ILI>
  </EQ_RELATION>
 </EQ_LINKS>
</WORD_MEANING>
```

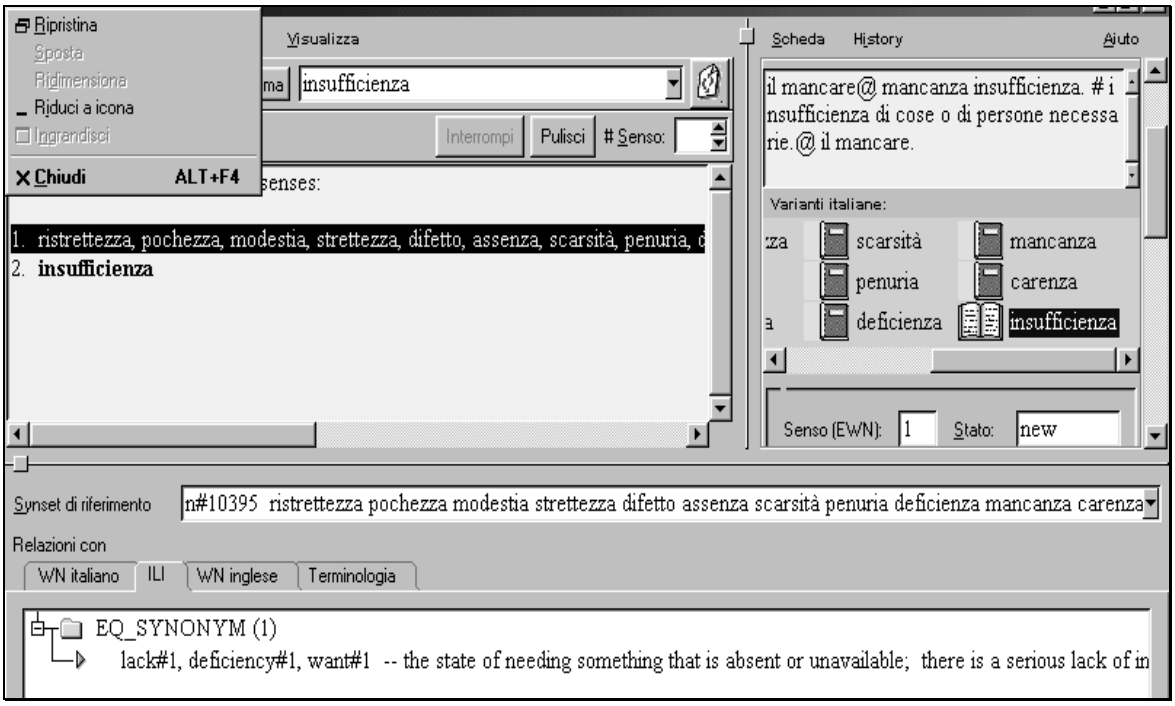


Fig. 8: Example of an IWN entry: "insufficienza"

3.2.3.4.2 Verbs

1. aumentare (to increase)

```
<WORD_MEANING ID="v@2298@" PART_OF_SPEECH="v">
 <VARIANTS>
  <LITERAL LEMMA="aumentare" SENSE="2" DEFINITION="diventare più grande più intenso o più numeroso."
STATUS="new"> </LITERAL>
  <LITERAL LEMMA="ingrandirsi" SENSE="1" STATUS="new"> </LITERAL>
  <LITERAL LEMMA="crescere" SENSE="2" STATUS="new"> </LITERAL>
  <LITERAL LEMMA="salire" SENSE="4" STATUS="new"> </LITERAL>
  <LITERAL LEMMA="accentuarsi" SENSE="1" DEFINITION="diventare più accentuato." STATUS="new">
</LITERAL>
 </VARIANTS>
 <INTERNAL_LINKS>
  <RELATION R_TYPE="has_hyperonym" ID="IR144856">
   <TARGET_CONCEPT ID="v@1640@" PART_OF_SPEECH="v">
    <LITERAL LEMMA="diventare" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  <RELATION R_TYPE="has_hyponym" ID="IR144857">
   <TARGET_CONCEPT ID="v@2336@" PART_OF_SPEECH="v">
    <LITERAL LEMMA="gonfiarsi" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  <RELATION R_TYPE="has_hyponym" ID="IR144858">
   <TARGET_CONCEPT ID="v@2337@" PART_OF_SPEECH="v">
    <LITERAL LEMMA="ricrescere" SENSE="2"> </LITERAL>
   </TARGET_CONCEPT>
```

```
  </RELATION>
  <RELATION R_TYPE="has_hyponym" ID="IR144859">
   <TARGET_CONCEPT ID="v@2338@" PART_OF_SPEECH="v">
    <LITERAL LEMMA="rinfrescare" SENSE="2"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  ……..

  </RELATION>
 </INTERNAL_LINKS>
<EQ_LINKS>
  <EQ_RELATION R_TYPE="eq_synonym" ID="ER144869">
   <TARGET_ILI ID="ILI144870" PART_OF_SPEECH="v" WORDNET_OFFSET="00093597"> </TARGET_ILI>
  </EQ_RELATION>
  <EQ_RELATION R_TYPE="eq_generalization" ID="ER144871">
   <TARGET_ILI ID="ILI144872" PART_OF_SPEECH="v" ADD_ON_ID="5502"> </TARGET_ILI>
  </EQ_RELATION>
  <EQ_RELATION R_TYPE="eq_generalization" ID="ER144873">
   <TARGET_ILI ID="ILI144874" PART_OF_SPEECH="v" ADD_ON_ID="5527"> </TARGET_ILI>
  </EQ_RELATION>
 </EQ_LINKS>
</WORD_MEANING>
```

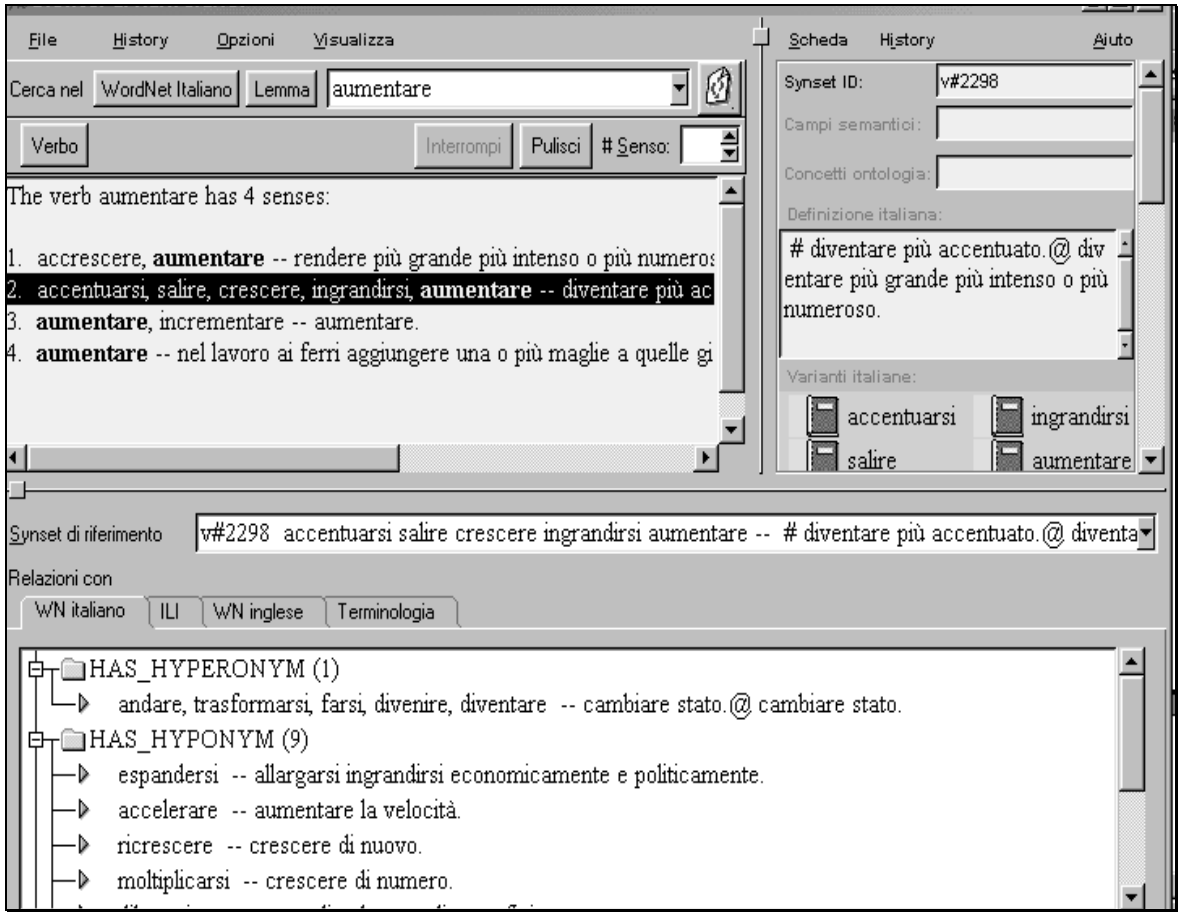


Fig. 9: Example of an IWN entry: “aumentare”

3.2.3.4.3 Adjectives

Abietto, spregevole, vile (abject)

```
<WORD_MEANING ID="a@2@" PART_OF_SPEECH="a">
 <VARIANTS>
  <LITERAL LEMMA="abietto" SENSE="1" DEFINITION="Che è spregevole vile"> </LITERAL>
  <LITERAL LEMMA="spregevole" SENSE="2" DEFINITION="Che è abietto"> </LITERAL>
  <LITERAL LEMMA="vile" SENSE="2" DEFINITION="Che  spregevole ignobile"
EXAMPLES="'Un'azione vile/spregevole/ignobile/abietta'"> </LI
TERAL>
 </VARIANTS>
 <INTERNAL_LINKS>
  <RELATION R_TYPE="xpos_near_synonym" ID="IR199203">
   <TARGET_CONCEPT ID="n@15259@" PART_OF_SPEECH="n">
    <LITERAL LEMMA="verme" SENSE="2"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  <RELATION R_TYPE="xpos_near_synonym" ID="IR199204">
   <TARGET_CONCEPT ID="n@19858@" PART_OF_SPEECH="n">
    <LITERAL LEMMA="abiezione" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  <RELATION R_TYPE="xpos_near_synonym" ID="IR199205">
   <TARGET_CONCEPT ID="n@20629@" PART_OF_SPEECH="n">
    <LITERAL LEMMA="viltà" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
  <RELATION R_TYPE="near_antonym" ID="IR199206">
   <TARGET_CONCEPT ID="a@42813@" PART_OF_SPEECH="a">
    <LITERAL LEMMA="ammirevole" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
 </INTERNAL_LINKS>
 <EQ_LINKS>
  <EQ_RELATION R_TYPE="eq_synonym" ID="ER199207">
   <TARGET_ILI ID="ILI199208" PART_OF_SPEECH="a" WORDNET_OFFSET="00673492">
</TARGET_ILI>
  </EQ_RELATION>
 </EQ_LINKS>
</WORD_MEANING>
```

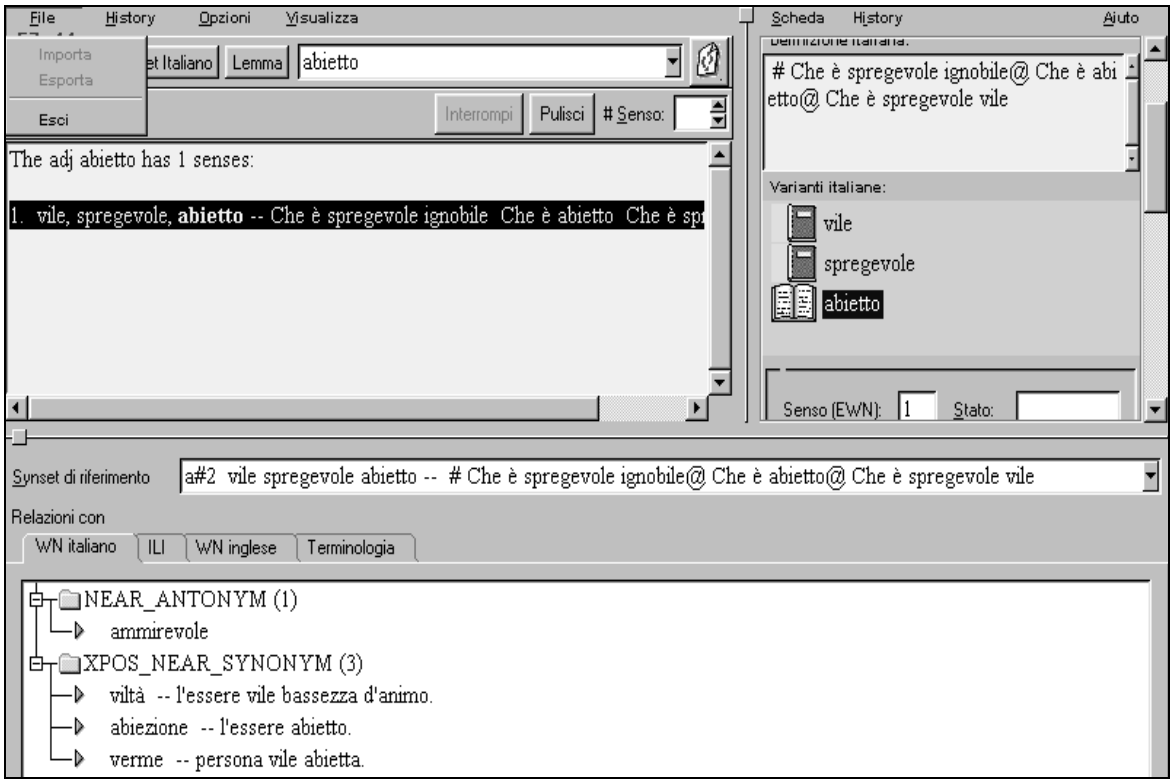


Fig.10: Example of an IWN entry: "abietto"

3.2.3.4.4 Instances

Capri (the island)

```
<WORD_INSTANCE ID="p@41457@" PART_OF_SPEECH="p">
 <VARIANTS>
  <LITERAL LEMMA="Capri" SENSE="1" STATUS="new"> </LITERAL>
 </VARIANTS>
 <INTERNAL_LINKS>
  <RELATION R_TYPE="belongs_to_class" ID="IR205444">
   <TARGET_CONCEPT ID="n@18567@" PART_OF_SPEECH="n">
    <LITERAL LEMMA="isola" SENSE="1"> </LITERAL>
   </TARGET_CONCEPT>
  </RELATION>
 </INTERNAL_LINKS>
</WORD_INSTANCE>
```

Fig. 11: Example of an IWN entry: "Capri"

3.2.3.5 Synoptic table of information types in the EWN and IWN lexicons.

The following table is a means to give an overview of the content of the typical IWN entry: for each entry component, it says whether the type of information is present or not and in which field of the record you one find it.

It is important to note that the table has the function to describe the potentiality of the linguistic model and that much of the information that is possible to express is only optional and it has not actually been massively codified in the data (only synonymy and hyp(er)onymy are not optional).

The information in E(/I)WN is not bilingual in the proper sense, since it is realized by means of an interlingual module and not via a direct cross-lingual bilingual link.In this sense, the "target language" notion is ambiguous: the target language is firstly the language of the query target wordnet, secondly it is the American-English of the ILI.The relation with the other wordnets is only established indirectly: each site maps its synset directly to the ILI and the lexicographer has no idea about how the translation will be realized in the various languages; the equivalence is possible only if the other wordnets link meanings to the same ILI-record.

The outcome of the translation among parallel wordnets depends on how the link to the ILI is realized (Peters et al., 1998).

If the same concept is present in the languages A and B but not in the ILI, the translation is not going to take place.

If the concept is realized in the same way in the languages A and B (with the same semantic-syntactic structure, with the same group of synonyms etc.) but it doesn't exactly match with an ILI synset, then the relation between the word meaning (wm)(A) and wm(B) will not be a equivalent synonymy.

During EWN, this problem has been studied through an analysis of the ILI gaps. Some recurrent gaps, due to different lexicalisation patterns in the various languages, have been highlighted and a model of a condensed and universal index of meaning was proposed (for further details, see Vossen at al., 2000).

The result of this work could be very important to better express the full potentiality of this resource in multilingual applications. The advantages of an **interlingual** rather than a **cross-lingual** approach in CLTR are discussed in Golzalo et al., 1998.

Table 11: Lexical Information in the EWN(/IWN) lexicon

	Entry component		**Present**	**Representation in the Lexicon**
1	headword		✓	LITERAL
2	Phonetic transcription			
3	variant form		✓	LITERAL+VARIANTS
4	inflected form			
5	Cross-reference			
6	**Morphosyntactic Information**			
	a	Part-of-speech marker	✓	PART_OF_SPEECH
	b	Inflectional class		
	c	Derivation	✓	RELATION R_TYPE “Derivation”
	d	Gender	✓	FEATURES
	e	Number	✓	FEATURES
	f	Mass vs. Count		
	g	Gradation		
7	Subdivision counter			
8	Entry subdivision		✓	The subdivision of each entry in different literals
9	Sense indicator		✓	SENSE
10	linguistic label		✓	USAGE
11	**Syntactic Information**			
	a	Subcategorization frame		
	b	Obligatority of complements		
	c	Auxiliary		
	d	Light or support verb construction		
	e	Periphrastic constructions		

	f	Phrasal verbs		
	g	Collocator		
	h	Alternations		
12	**Semantic Information**			
	a	Semantic type	✓	« has_hyperonym » relation and, by means of the Base Concepts set, Ontological information
	b	Argument structure		
	c	Semantic relations	✓	Internal relations
	d	Regular polysemy	✓	Multiple inheritance with disjunction and conjunction features
	e	Domain	✓	Sublanguage and information in the Domain Ontology.
	f	Decomposition		
13	Translation		✓	TL equivalent reached via an equivalent relation
14	Gloss		✓	DEFINITION
15	Near-equivalent		✓	RELATION R_TYPE “near_synonym”
16	Example phrase (straightforward)		✓	EXAMPLE
17	Example phrase (problematic)		✓	EXAMPLE
18	multiword unit		✓	LITERAL
19	Subheadword *also* secondary headword			
20	usage note		✓	USAGE LABEL
21	Frequency			

3.2.4 PAROLE-SIMPLE lexicons

3.2.4.1 General overview of the PAROLE-SIMPLE lexicons

SIMPLE is a project sponsored by EC DGXIII in the framework of the Language Engineering programme. This project - which has ended on April 30th 2000 - has developed core semantic lexicons for 12 languages (Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish), with a harmonised common model that encodes structured "semantic types" and semantic (subcategorisation) frames.

SIMPLE should be considered as a follow up to the PAROLE project, because it adds a semantic layer to a subset of the existing morphological and syntactic layers developed by PAROLE. The semantic lexicons (about 10,000 word meanings) have been built in a harmonised way for the 12 PAROLE languages. Both are based on EAGLES recommendations. These lexicons are partially corpus-based, exploiting the harmonised and representative corpora built within PAROLE. The lexicons have been designed bearing in mind a future cross-language linking: they share and are built around the same core ontology and the same set of semantic templates. The "base concepts" identified by EuroWordNet (about 800 senses at a high level in the taxonomy) are used as a common set of senses, so that a cross-language link for all the 12 languages is already provided automatically through their link to the EuroWordNet Interlingual Index (see http://www.let.uva.nl/~ewn).

The PAROLE-SIMPLE Lexicons (henceforth P-S) are three-layered lexicons, whose entries are encoded at the *morphological*, *syntactic* and *semantic* level:

- The PAROLE part of P-S contains ~20.000 entries (verbs, nouns, adjectives, numerals, adverbs, pronouns, prepositions, conjunctions, determiners, interjections), each encoded at the morphological and syntactic level

- The SIMPLE part of P-S contains ~10.000 senses of PAROLE entries (~7000 nouns, ~2000 verbs and ~1000 adjectives), each linked to the relevant syntactic descriptions

Although PAROLE and SIMPLE respectively correspond to a morphosyntactic and a semantic lexicon, they should regarded as a unique and coherent body, since they have been both built in accordance to the GENELEX relational model. Moreover the three layers are interlinked, so that, for instance, argument positions defined at the semantic layer in SIMPLE are associated to the relevant syntactic positions defined in the PAROLE lexicon, and complex interactions between syntactic alternations and semantic interpretations can be represented. Each piece of linguistic information is encoded by means of SGML tags, defined in the GENELEX PAROLE-DTD. P-S lexicons do not contain multiword expressions.

The P-S lexicons are publicly available through ELRA. Samples of the PAROLE-SIMPLE entries for the 12 lexicons are available at the project Web site: http://www.ub.es/gilcub/SIMPLE/simple.html

In what follows, we give a brief description of the syntactic part of P-S, to then pass to discuss in more details the linguistic model underlying the semantic encoding in SIMPLE and the organization of the semantic entries.

3.2.4.2 The morphosyntactic layer (PAROLE)

The following is the morphosyntactic information represented in the P-S lexicons. Each piece of information corresponds to specific SGML elements or attributes, as defined by the PAROLE-DTD:

***Morphological Level*:**

- Grammatical category and subcategory
- Gender, number, person, mood
- Inflectional class
- Modifications of the lemma

***Syntactic Level*:**

- **Idiosyncractic properties of an entry wrt a given syntactic construction:**
  - Idiosyncratic behaviour with respect to specific syntactic rules (passivisation, middle, etc.)
  - Subclass; auxiliary (*only for verbs*)
  - Mass/count, 'pluralia tantum' (*only for nouns*)
  - Attributive vs. predicative function, gradability (*only for adjectives*)
  - Semantic subtype and part of speech to which they are related (*only for adverbs*)
- **Subcategorization frames:**
  - List of syntactic positions (at most 4: P0, P1, P2, P3)
  - Optionality of a position
  - Syntactic constraints and property of the possible 'slot filler'
  - Grammatical function (*for verbs and deverbal nouns*)
  - Possible syntactic realizations of the position
  - Morphosyntactic and/or lexical features (agreement, prepositions and particles introducing clausal complements)

- Information on control (subject control, object control, etc.) and raising properties
- Position of the lemma with respect to its complements

3.2.4.3 The semantic layer (SIMPLE)

The SIMPLE model is based on the recommendations of the EAGLES Lexicon/Semantics Working Group (http://www.ilc.pi.cnr.it/EAGLES96/rep2) and on extensions of Generative Lexicon theory. An essential characteristic is its ability to capture the various dimensions of word meaning. The basic vocabulary relies on an extension of "qualia structure" (cf. Pustejovsky 1995) for structuring the semantic/conceptual types as a representational device for expressing the multi-dimensional aspect of word meaning.

SIMPLE also provides a common "library" of language independent **templates**, which act as "blueprints" for any given type - reflecting the conditions of well-formedness and providing constraints for lexical items belonging to that type.

The SIMPLE model thus contains three types of formal entities (cf. also fig. 12):

1. **SemU** - word senses are encoded as *Semantic Units* or *SemU*. Each SemU is assigned a *semantic type* in the ontology plus other sorts of information which are intended to identify a word sense, and to discriminate it from the other senses of the same lexical item. SemUs are language specific. SemUs which identify the same sense in different languages will be assigned the same semantic type.

2. **(Semantic) Type** - it corresponds to the semantic type which is assigned to SemUs. Each type involves, among others, structured information, organized in the four Qualia Roles, adopted in the Generative Lexicon framework. The Qualia information is sorted out into *type-defining information* and *additional information*. The former is information which intrinsically defines a semantic type as it is. In other words, a SemU can not be assigned a certain type, unless its semantic content includes the information that defines that type. On the other hand, additional information specifies further semantic components a SemU, rather than entering into the characterization of its semantic type.

3. **Template** - a schematic structure which the lexicographer uses to encode a given lexical item. The template expresses the semantic type, plus additional information, e.g. domain, semantic class, gloss, predicative representation, argument structure, polysemous classes, etc. Templates are intended to guide, harmonize, and facilitate the lexicographic work. A set of top templates have been prepared during the specification phase, while more specific ones may be eventually elaborated by the different partners according to the need of encoding more specific concepts in a given language.

The SIMPLE model provides the formal specification for the representation and encoding of the following information (the items marked with an asterisk, refer to the information which is obligatorily encoded for every word sense):

- Semantic type (*)
- Domain information (*)

- Glossa (*)
- Argument structure (*)
- Semantic roles and selectional restrictions on the arguments (*)
- Event type for verbs (*), to characterize their actionality behaviour
- Link of the arguments to the syntactic subcategorization frames, as represented in the PAROLE lexicons (*)
- Type hierarchy information
- Qualia information, in terms of both features and relations between SemUs
- Information about regular polisemous alternation in which a word sense may enter
- Information concerning cross-part of speech relations (e.g. "intelligent" - "intelligence"; "writer" - "to write")
- Eventual collocations from the corpus
- Synonymy relations

The hierarchy of types has been further subdivided in three layers (for a sample see fig. 13 below):

- ***The Core Ontology*** - it is formed by those types which have been identified as the central and common ones for the construction of the different lexicons in SIMPLE. The Core Ontology has been elaborated according to the following criteria:

  1. Their central position in the organization of the lexicon;
  2. The fact that they are widely acknowledged in the linguistic, NLP literature and in applied systems as core notions for the semantic characterization of words;
  3. The low level of granularity of the semantic description they provide, which also ensures their multilingual usability. Therefore, the elements of the Core Ontology represent the highest nodes in the hierarchy of types.

- ***Recommended Ontology*** - this is formed by more specific types (lower nodes in the hierarchy), which provide a more granular organization of the word-senses.
- ***(Language) Specific types*** - more detailed types may be created in order to organize a lexicon for language-, domain- or application-specific needs. These types are not provided in the specification phase, and can be eventually added if their elaboration is consistent with the organization of the rest of the SIMPLE model.

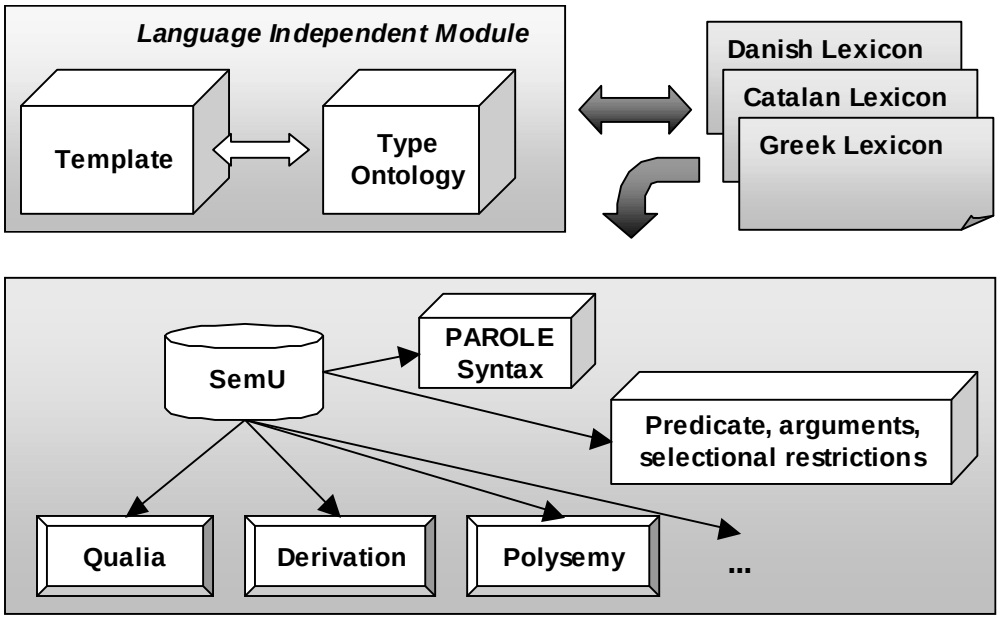


Fig. 12: SIMPLE overall structure

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]

Fig. 13: The SIMPLE ontology: a sample

Presently, the SIMPLE semantic lexicons do not contain any multilingual information nor any multilingual link, although the 12 lexicons have been developed in parallel and according to a unique, highly constrained linguistic model. The turning of SIMPLE lexicons into real multilingual resources is envisaged in the near future, and experiments in this sense are already ongoing (cf. for instance Villegas *et al.*, 2000).

Word senses to be encoded for each lexical head have been usually identified by using medium-size monolingual dictionary. As a general constraint, all the senses belong to PAROLE entries. SIMPLE lexicons are *general purpose lexicons* and the lexicon population has been determined according to the two following criteria:

1. guaranteeing that all the semantic types of the ontology are instantiated (so that different semantic areas of the lexicon are represented);

2. closure of the entries, so that every sense of a given PAROLE entry have been encoded.

The information contained in the SemU has been selected on two basis: (i.) information provided by medium-size resources (either manually or automatically extracted) (ii.) corpus-evidence. The various lexicons differ depending on the balance between these two strategies.

Semantic information describing the SemU content is represented in terms of three formal entities specified by the SGML DTD:

1. *Features* - domain information, semantic class, template type, etc.

2. *Semantic relations between SemUs* - They include (i.) 4 hierarchical organized sets of Qualia relations (one for Quale); (ii.) derivational relations; (iii.) polysemous information; (iv.) synonymy; (v.) collocations

3. *Predicative Representation* - specifies the predicate to which a SemU is associated. On turn a predicate is specified by the number of its arguments, semantic roles, selectional preferences on the arguments.

The following is a small sample of the 66 semantic relations adopted in SIMPLE:

Name	Description	Example	Type
Is_a_member_of	<$SemU_1$> is a member or element of <$SemU_2$>.	<senator>;<senate>	Constitutive
Is_a_part_of	<$SemU_1$> is a part of <$SemU_2$>	<head>;<body>	Constitutive
Used_for	<$SemU_1$> is typically used for <$SemU_2$>	<eye>;<see>	Telic
Used_as	<$SemU_1$> is typically used with the function which is expressed by <$SemU_2$>	<wood>;<material>	Instrument

Resulting_state	<$SemU_1$> is a transition and <$SemU_2$> is the resulting state of the transition	<die>;<dead>	Constitutive
Created_by	<$SemU_1$> is obtained, or created by a certain human process or action <$SemU_2$>	<book>;<write>	Artifactual_agentive
Purpose	<$SemU_2$> is an event corresponding to the intended purpose of <$SemU_1$>	<send>;<receive>	Telic

3.2.4.4 The structure of an entry in the PAROLE-SIMPLE lexicons

3.2.4.4.1 Morphological level

```
<MuS
            id="MUS_aumentare_VERB"  %% morphological unit identifier%%
            gramcat="VERB"
            autonomy="YES"
            synulist="SYNU_aumentare_V SYNU_aumentare_V_2">  %%link to the syntactic units
describing the syntactic behavior of the entry%%
            <Gmu
                 inp="GINP_294">  %%inflectional code%%
                 <Spelling>aumentare</Spelling></Gmu></MuS>
```

3.2.4.4.2 Syntactic level

```
<SynU
                id="SYNU_aumentare_V"  %%syntactic unit identifier%%
                 naming="aumentare"
                example="Il pane aumenta di dieci lire"
                comment="inadj"
                description="i-adj_ppdi*)-xe">  %%syntactic description identifier%%

             <CorrespSynUSemU   %%link to the semantic units%%
                   targetsemu="USem3981"
                   correspondence="ISObivalent"></SynU>


<SynU
                id="SYNU_aumentare_V_2"
                naming="aumentare"
                example="aumentare i prezzi del 10 per cento"
                comment="tr  P2 tr/P1 in"
                description="t-adj_ppdi*)-xa">
             <CorrespSynUSemU
                   targetsemu="USem3980"
```

```
                    correspondence="ISOtrivalent"></SynU></SynU>
```

%%Every SynU describes a particular syntactic behavior of the morphological unit in the Description object. This on turn speciefies the Self object (describing the property of the entry in the given syntactic context), and the Construction object (specifying the subcategorization frame associated to the given syntactic description).%%

```
<Description
                id="i-adj_ppdi*)-xe"
                example="Il pane aumenta di dieci lire"
                self="SELF_V_xe"
                construction="i-adj_ppdi*)">

<Self
                   id="SELF_V_xe"
                   intervconst="I_V_xe">

<IntervConst
                   id="I_V_xe"
                   syntagmatl="S_T_V_xe">
<SyntagmaT
                   id="S_T_V_xe"
                   syntlabel="V"
                   featurel="T_AUX_essere">
                   <SyntFeatureClosed
                         featurename="MORPHSUBCAT"
                         value="MAIN"></SyntagmaT>

<AuxFeature
                   id="T_AUX_essere"
                   value="essere">
```

%%The construction shown below describes the intransitive reading of the verb 'aumentare' (to increase), with two syntactic positions%%

```
<Construction
                id="i-adj_ppdi*)"
                syntlabel="Clause"
                selfinsertion="1">
                     <InstantiatedPositionC
                          range="0"
                          optional="YESO"
                          positionc="P_subj">
                     <InstantiatedPositionC
                          range="1"
                          optional="YESO"
                          positionc="P_adj_ppdi*"></Construction>
```

%%The object 'PositionC' describes the grammatical function and the realization of a syntactic position%%

```
<PositionC
                id="P_subj"
                function="SUBJECT"
                syntagmacl="S_NT_np">

<PositionC
                id="P_adj_ppdi*"
                function="ADVERBIAL"
                syntagmacl="S_NT_ppdi3">
<SyntagmaNTC
                id="S_NT_np"
                syntlabel="NP"></SyntagmaNTC>
<SyntagmaNTC
                id="S_NT_ppdi3"
```

```
                syntlabel="PP"
                featurel="T_di">
                <SyntFeatureClosed
                     featurename="SYNSUBCAT"
                     value="WITHOUTDET"></SyntagmaNTC>

<LexFeature
                   id="T_di"
                   featurename="INTROD"
                   value="di"
                   mu="MUS_di">
```

3.2.4.4.3 Semantic level

%%The following SemU describes the inchoative meaning of the verb 'aumentare' (to increase)%%

```
<SemU
     id="USem3981"
     naming="aumentare"
     example="la popolazione è aumentata del 10 %"
     comment="BC 10"
     freedefinition="accrescersi, salire di prezzo"
     weightvalsemfeaturel="TSVP_CHANGE_TS_classificateur_de_verbe_C
WVSFDirectionUpPROT WVSFEventTypeTransitionPROT WVSFTemplateChangeofvaluePROT
WVSFUnificationPathRelationalchange-AgentivePROT">   %%These features describe the semantic type, the position of this type in the overall ontology, the event type%%
<PredicativeRepresentation
          typeoflink="Master"
          predicate="PREDaumentare-2"> %%Name of the predicate to which the given SemU is associated, and type of the association%%
     <RWeightValSemU
          weight="PROTOTYPICAL"
          comment="cambiare"
          target="USem3939"
          semr="SRIsa"> %%Semantic relation. The example reports a case of Is_a link%%
      <RWeightValSemU
          weight="ESSENTIAL"
          comment="aumentare"
          target="USem3980"
          semr="SRPolysemyChangeofvalue-Causechangeofvalue"> %%Semantic relation expressing a regular polysemous link withy the SemU corresponding to the causative reading of the same verb%%
       <RWeightValSemU
          weight="PROTOTYPICAL"
          comment="DUMMYmaggioreA1"
          target="USemD5448"
          semr="SRResultingstate">
     <RWeightValSemU
          weight="PROTOTYPICAL"
          comment="cambiamento"
          target="USem3960"
          semr="SRAgentive">
<Predicate
     id="PREDaumentare-2"
     naming="aumentare-2"
     type="LEXICAL"
     multilingual="No"
     argumentl="ARG0aumentare-2 ARG1aumentare-2">   %%Number of semantic arguments of the predicate associated to the SemU via the predicative representation%%
```

```
<Argument
     id="ARG0aumentare-2"
     semanticrolel="RoleProtoPatient"
     informargl="INFARGT90"> %%Semantic role of the argument%%
<Argument
     id="ARG1aumentare-2"
     semanticrolel="RoleUnderspecified"
     informargl="INFARGT96">
<InformArg
     id="INFARGT90"
     weightvalsemfeaturel="WVSFTemplateEntityPROT"> %%Selectional preferences on the
arguments%%
<InformArg
     id="INFARGT96"
     weightvalsemfeaturel="WVSFTemplateAmountPROT">
```

3.2.4.5 Synoptic table of information types in the PAROLE-Simple lexicons.

In the following tables, we give an overview of the content of the dictionaries investigated in this survey on the basis of the "Lexical Information in bilingual resources" grid.

1. *Entry component* - name of the relevant component of the lexicon
2. *Present* - it marks whether a component is represented in P-S
3. *Representation in P-S* - it says where and how the component is represented in the P-S lexicons.

Table 12: Lexical Information in the PAROLE-SIMPLE lexicons

	Entry component	**Present**	**Representation in P-S**
1	Headword	✓	It is the value of the `id` attribute in the Morphological unit
2	Phonetic transcription[4]		
3	Variant form		
4	Inflected form	✓	Morphological units contain a link to the inflectional tables where number, gender, mood,

[4] The phonetic transcription will be encoded in the continuation of the PAOLE-Simple project, the Italian National Project CLIPS.

				tense information is contained, as well as the particular way in which the lexeme is inflected
5	Cross-reference			
6	**Morphosyntactic information**			
	a	Part-of-speech marker	✓	Value of the gramcat attribute in the Morphological unit
	b	Inflectional class	✓	Morphological units contain a link to the inflectional tables where number, gender, mood, tense information is contained, as well as the particular forms of a given entry
	c	Derivation	✓	Cross part of speech relations are marked through derivational semantic relations between SemUs
	d	Gender	✓	Expressed in the Ginp associated to a Morphological Unit
	e	Number	✓	Expressed in the Ginp associated to a Morphological Unit
	f	Mass vs. Count	✓	Expressed in the Morphological Unit
	g	Gradation	✓	Expressed in the Morphological Unit
7	Subdivision counter			
8	Entry subdivision		✓	Value of the attribute id in the `SemU` object
9	Sense indicator		✓	This information is captured by the values of the attributes `naming`, `example` and `comment`, which conjointly give clues to show the specific sense encoded in the `SemU`
10	Linguistic label		✓	Only for information about the terminological domain
11	**Syntactic information**			
	a	Subcategorization frame	✓	Described in the Syntactic Units specifying the number of positions, the syntactic realization (type of phrase, introducer, etc.). Each syntactic description is then linked to a Semantic Unit, and the argument structures are linked to their syntactic realizations
	b	Obligatority of complements	✓	Marked in the Syntactic Unit
	c	Auxiliary	✓	Marked in the Self object associated to a Syntactic Unit
	d	Light or support verb construction		
	e	Periphrastic constructions		
	f	Phrasal verbs	✓	
	g	Collocator	✓	Optionally encoded in the semantic layer: typical subject, typical object, etc.
	h	Alternations	✓	Represented in terms of syntactic descriptions (i.e. subcategorization structures) linked in a Frameset
12	**Semantic information**			

	a	Semantic Type	✓	Represented as link between a Semantic Unit and a node in the Ontology of semantic types
	b	Argument Structure	✓	Represented in the Predicative Representation associated to Semantic Units: it contains a link between the Semantic Unit and a predicate, on turn defined in terms of the number of arguments, their thematic roles, and selectional preferences
	c	Semantic relations	✓	Represented as relations between Semantic Units (e.g. hyperonymy, meronymy, and many others)
	d	Regular polysemy	✓	Represented as relations between Semantic Units
	e	Domain	✓	Represented as link between a Semantic Unit and a node in a hierarchy of domains
	f	Decomposition		
13	translation			
14	gloss		✓	In the attribute `freedefinition` a gloss is specified, as derived from a medium-sized monolingual dictionary
15	Near-equivalent			
16	Example phrase (straightforward)		✓	This is the value of the attribute `example`
17	Example phrase (problematic)			
18	multiword unit			
19	subheadword (secondary headword)			
20	usage note			
21	frequency			

The morphosyntactic, syntactic, and semantic information represented in P-S lexicons can be combined to carry out various types of tasks. In what follows, we will illustrate how the P-S lexical entries can be used to handle some of the cross-lingual lexical phenomena selected for the lexicon survey task in the ISLE project. There are two major caveats to consider:

1. As already noticed above, cross-lingual links are not explicitly part of the P-S lexicons. Hence, what is given here should be better regarded as an illustration of possible ways to tackle some cross-lingual lexical phenomena, given the information available in P-S and the architecture of these lexicons;

2. Multiwords expressions are not currently represented in P-S (they are added in a few extensions within National Projects).

3.3 Resources for MT systems

3.3.1 Eurotra Bilingual Lexical Resources

Eurotra was a transfer based and syntax driven MT system which dealt with 9 languages (Danish, Dutch, German, Greek, English, French, Italian, Spanish and Portuguese). Monolingual and bilingual lexical resources were developed for all languages and, in the case of bilingual, in all possible directions. All information was encoded as Feature-Value pairs in ASCII files. Eurotra is no longer developed nor supported (although there are MT systems closely related, such as PaTrans), but the interest in considering its lexical resources comes from the efforts made to minimize the transfer components by agreeing in the information to be dealt with for translating among the 9 languages.

Transfer was performed in EUROTRA between the Interface Structure of a source language and the Interface Structure of a target language. The strategy adopted in the EUROTRA Translation System with respect to transfer is to start from Interface Structure representations which overcome, as much as possible, structural differences between languages. This is done by treating some phenomena interlingually (like semantic treatment of tense and aspect) and by neutralising different surface realisations (as, for example, elevating prepositions of governed elements, defining common argument structure definitions, etc.). This strategy aims at keeping transfer as simple as possible by reducing its operations, in the best case, to the copying of interlingual information and neutralised structures.

Thus, sense distinctions were to be identified in monolingual analysis, and the bilingual resources refer to these sense distinctions for relating two lexical entries as translational equivalent. Information that is used to distinguish different readings mostly concerns to argument structure differences, semantic typing of heads, and semantic typing of the arguments. Terminological readings were also taken into account.

3.3.1.1 Bilingual Information in an Eurotra entry

Table 13: Summary of the information types in the Eurotra lexicons

	Entry component	Present	Information content
1	headword	✓	lexical unit (lu): lemma
2	phonetic transcription		
3	variant form	✓	alternative spellings were encoded as different lexical units
4	inflected form		NO, but information was used when needed in the form of attribute-value features
5	Cross-reference		
6	**Morphosyntactic information**		

	a	Part-of-speech marker	✓	
	b	Inflectional class	✓	
	c	Derivation	✓	Major derivational patterns encoded as features
	d	Gender	✓	Encoded as a feature (gen)
	e	Number	✓	Encoded as a feature (nb)
	f	Mass vs. Count	✓	Encoded as semantic typing (sem)
	g	Gradation		
7	Subdivision counter			
8	Entry subdivision			
9	Sense indicator			
10	linguistic label			
11	**Syntactic information**			
	a	Subcategorization frame	✓	Exhaustive subcategorizaton information in terms of syntactic complements and arguments related.
	b	Obligatority of complements	✓	Included in subcategorization information
	c	Auxiliary	✓	Encoded for those languages which required it as a feature
	d	Light or support verb construction	✓	Support verbs constructions were encoded in predicative nouns, where the different verbs chosen by the particular noun are encoded as values of different features (see below)
	e	Periphrastic constructions		
	f	Phrasal verbs	✓	Phrasal verbs identified during analysis become a lexical unit
	g	Collocator		
	h	Alternations	✓	Encoded in subcategorization information
12	**Semantic information**			
	a	Semantic type	✓	Semantic typing but different systems used for different languages
	b	Argument structure	✓	Argument structure and the semantic typing of the arguments were encoded for the major categories
	c	Semantic relations		
	d	Regular polysemy		
	e	Domain	✓	Terminological items where marked as such but no domain classification
	f	Decomposition		

13	Translation	✓	Encoded in the bilingual transfer modules
14	Gloss		
15	Near-equivalent		
16	Example phrase (straightforward)		
17	Example phrase (problematic)		
18	multiword unit	✓	Different treatments. See Complex transfer below
19	subheadword *also* secondary headword		
20	usage note		
21	Frequency		

3.3.1.2 Simple Transfer

Transfer is simple when the lexical units of the source language are exchanged for the lexical units of the target language, and all other information contained in the structure and in the set of features is copied. It is complex if the structure is transformed and information contained in the features changed.

Simple transfer is performed mainly by the built-in default translation mechanism of all the translators in the system. The default translator copies structures and those features declared both in the source level and target level feature declaration. The only explicit operation we need is for simple lexical transfer, i.e. feature rules (f-rules) which change the lexical unit value from the source language into the target language lu-value. This component together with lexical monolingual information for both the SL and the TL can be considered bilingual dictionaries. To perform the mapping from a lexical entry in the SL onto one lexical entry in the TL the lu-value has to be specified with the reading number (an attribute-value pair which identifies sense distinctions based on formal differences in the encoding of the entries), when more than one reading of a lexical unit exists (3.2.1.2.3).

Lexical Disambiguation is performed through the same rules that perform lexical transfer if the relevant disambiguating feature is present at the leaf node.

3.3.1.3 Complex Transfer

For complex transfer, explicit feature and structural rules which overwrite the built-in default translator where used (see section 5).

Complex Lexical Transfer

We have already said that the transfer translator has one main function in the Eurotra Translation System, namely to perform simple lexical transfer between two languages, to map lexical units from one language onto lexical units of another. This is not always possible through simple lexical transfer rules, which perform one-to-one mappings. There are two cases of special relevance. First, when it is required to express the context of a lexical unit to decide the right translation. Second, where there is no one-to-one mapping.

Disambiguation through context

In order to contextualise a lexical unit the mother and/or sister nodes have to be described. This is done by means of structure rules, which do not delete information, or change structure, but perform the translation of a lexical unit in a given context.

example:

seit => desde_hace
=> desde

The strategy here is to specify the contexts which determine the translation of the preposition 'seit' into either 'desde' or 'desde_hace'.

```
tseit = PP:{cat=pp}
           [P:{cat=p,d_lu=seit},
           (ADVP:{cat=adv};
            NP:{cat=np}
              [N:{},
              ^(AP:{cat=ap,d_semtype=temp};
               ORD:{cat=ordp};
               DEM:{cat=demp};
               N2:{cat=np,dtype=poss})])
           ]
 =>
        PP<P:{e_lu=desde},NP<N,AP,ORD,DEM,N2>>.


tseit2 = PP:{cat=pp}
           [P:{cat=p,d_lu=seit},
           NP:{cat=np,d_msdefs~=msdef}
              [N:{cat=n,d_semtype=temp},
               CAR:^{cat=cardp},
               AP:^{cat=ap,d_semtype~=temp},
               QUANT:^{cat=quantp}]]
```

```
=>
        PP<{cat=p,e_lu=desde_hace},NP<N,CAR,AP,QUANT>>.
```

For other cases of lexical collocations the context the lexical unit stands in has to be fully specified. This is the case for the German verb 'kommen' which in the context of 'zum Einsatz kommen' is translated for 'entrar en funcionamiento'.

```
teinsatzkommen = S:{cat=s}[V:{cat=v,d_lu=kommen},
                           ARG1:{role=arg1},
                           ~:{cat=pp}
                              [~:{cat=p,d_lu=zu},
                               ~:{cat=np,cs=CS,argtype=AT}
                                  [~:{cat=n,d_lu=einsatz}]],
                           ANY:*{role=mod}]
=>
                S<V:{e_lu=entrar,e_isframe=arg1_2},
                  ARG1,

{cat=np,role=arg2,cs=CS,argtype=AT,e_msdefs=msabs,nb=sing}
                    <{e_lu=funcionamiento}>,
                  ANY>.
```

The German expression 'sich auf einer Umlaufbahn bewegen' has to be translated into Spanish as 'describir una órbita'.

```
tsichbewe = S:{cat=s}
              [V:{d_lu=sich_bewegen,isframe=arg2_PLACE},

~:{cat=np,role=arg2,nb=NB,argtype=AT,person=PE,cs=CS}
                  [N:{cat=n,d_lu=satellit},

MOD:{role=mod,cat=quantp}[M:{cat=quant,d_lu=all}],
                   MOD2:{role=mod,cat=advp}],
               ~:{cat=pp,role=argPLACE}
                  [~:{cat=p,d_lu=auf},

~:{cat=np,nb=NU,argtype=A,d_msdefs=MS,person=P,cs=C}
                      [N2:{cat=n,d_lu=umlaufbahn},
MO:*{role=mod}]],
               ANY:{role=mod}]
=>
       S:{dia=activ}<V:{e_lu=describir},
         {cat=np,role=arg1,nb=NB,argtype=AT,person=PE,cs=CS}
             <N,MOD<M,MOD2>>,

{cat=np,role=arg2,nb=NU,argtype=A,e_msdefs=MS,person=P,cs=C}
             <N2,MO>,
         ANY>.
```

No structural correspondence

The first case where the mapping is not one-to-one happens when one entry in the SL has to be mapped onto two entries of the TL.

The second case of no one-to-one mapping, and thus another important source of complex lexical transfer, occurs when lexical units of one language have no direct correspondence to lexical units in other languages. Lexical units may correspond to structure or may have no correspondence at all.

A very frequent case occurs with complex lexical units (compound units, fixed phrases and idioms) which have no correspondant lexical unit in the TL. These complex lexical units have to be transformed in transfer into more complex phrasal structures; which often involves category change. This is done by deleting the whole phrase at the left hand side of a structural rule and creating the new phrase at the right-hand side.

examples :

DE: in letzter Zeit	ES: últimamente
EL: aurio	ES: en el futuro
EN: confidently	ES: con seguridad
DE: beispielsweise	ES: por ejemplo
DA: genopbygningsperiode	ES: período de reconstrucción
DE: Kommunikationsweg	ES: enlace de comunicación
DE: Seekabel	ES: cable submarino

rules:

```
tult = ~:{cat=pp}
           [~:{cat=p,d_lu=in},
            ~:{cat=np}[~:{cat=n,d_lu=zeit},
                       ~:{cat=ap}[~:{cat=adj,d_lu=letzt}]]]
=>
      {cat=advp}<{cat=adv,e_lu='últimamente'}>.


tseekabel = ~:{cat=n,d_lu=seekabel,nb=N}
=>
             {cat=n,e_lu='cable',nb=N},{cat=ap,role=mod}

<{cat=adj,e_lu='submarino'}>
```

3.3.2 MT systems Metal and Logos

3.3.2.1 Transfer conditions

The following description of transfers is based on an investigation of the transfer possibilities of both the LOGOS and the METAL translation system. It maps their possibilities into a common framework proposal which will be used as a basis for the specification of the OLIF interchange format as developed in the OTELO project.

The description assumes that there is a syntactic tree as input of the transfer phase. This tree follows an X-bar scheme, and assumes a flat structure of the head and all its modifiers on the XP level; something like:

The head of the construction is marked.

All nodes are assumed to have features and values attached; these features and values cover syntactic functions (like subject, deep-direct object etc.) as well as (morpho)syntactic information (like part-of-speech, gender, etc.).

The idea is to select 1:n transfers by describing tests and actions. Tests and actions can be described as tree configurations. To give an example for a test:

de	en
bestehen <u>auf</u> etw.	*insist on sth*
bestehen <u>aus</u> etw.	*consist of sth*

So transfer is selected on the basis of the existence of a PP with a certain preposition. Depending on the value of the canonical form of this preposition, the transfer is selected. More examples can be found in (Thurmair, 1990).

Similarly, transfer actions would be described:

de	en
er gefällt mir	*I like him*

In this case, the grammatical functions must be changed: The German subject node will become the English indirect object node, and the German direct object will be assigned the English subject. This again can be expressed in terms of nodes and feature decoration.

In METAL, these phenomena are called complex lexical transfer, as it is not structural transfer (which should not involve lexical items) because it is triggered by lexical units, but it is also not just lexical replacement as it has effects on the syntactic structure and tree environment.

The following section presents an overview of the different possibilities and the features involved for METAL and LOGOS. It covers the majority of cases of complex lexical transfer, and it specifies the features and tree structures which are accessed in these operations.

While the current report describes the state of the art in transfer-based MT (looking at Globalink's rule editor in the Barcelona technology shows that the same mechanisms are used there), we should go beyond this level of description in ISLE. What these MT systems really do is a kind of word sense disambiguation at transfer time, and they try to find clues of which sense could have been meant in a given constellation. However, having more elaborated semantic machinery as proposed by SIMPLE could ease the task of transfer, by moving the sense disambiguation into the analysis part, and have an easier transfer part then. Even then, however, significant machinery is needed to describe collocational patterns, multiword expressions, and the like.

So we would still need the morphosyntactic machinery, but enrich it by information on semantic / pragmatic level.

The OTELO LDB will offer the user the option of specifying conditions for transfer relations. Since the statement and manipulation of these conditions often requires more extensive linguistic and system knowledge, only users with *administrator* access to the DB will have the authorization to create/modify them.

The OTELO definition should represent general linguistic requirements as reflected in the current specifications for relevant MT systems.

Parts-of-Speech for which Transfer Conditions can be Formulated

Transfer conditions should be definable for the following parts-of-speech:

- Noun
- Verb
- Adjective
- Adverb
- Preposition

Not all systems support conditions for all of the above five parts-of-speech, e.g., Logos does not permit users to generate Semtab rules indexed from prepositions.

⇒ Note: How to handle lexical information that is generated via OTELO, but richer in detail than an actual MT system allows, is a topic for further discussion.

Content of Conditions

Transfer conditions generally define a context for the translation of a source word/phrase into a target word/phrase. These conditions consist of:

a) The specification of context elements for the word/phrase. (These elements usually fall within the syntactic frame defined for that particular word/phrase.)

b) Tests on the features/values associated with these context elements.

The context elements are categorized based on their part-of-speech. Tests on context elements can be tests on feature values that are assigned in the lexicon, as well as feature values that are assigned in the analysis process.

⇒ Note: Whether we need to incorporate consistency checks to reconcile transfer conditions for a given word/phrase with its syntactic frame is open to discussion. This would in any case be difficult to do across the board, since some systems, e.g., Logos, do not have easily accessible codings for syntactic frames.

In addition to statements regarding context elements, a transfer condition can specify a test on the source word/phrase itself as a condition for translation.

Context Elements

The part-of-speech of a word/phrase determines the types of elements that can constitute a context for transfer. (1.) through (5) detail suggested context elements for the parts-of-speech listed in above.

1 Context Elements for Nouns

- Attached prep phrase(s) = **N** PP…
- Attached possessive phrase = **N** (of) N
- Descriptive adjective = Adj **N**
- Prep in phrase in which noun = Prep **N**
  Is object of prep

2 Context Elements for Verbs

- Noun arguments = **V** N(Subj), N(DO), N(IO)
- Attached prep phrase(s) = **V** PP…
- Adverb = **V** Adv
- Predicate adjective = **V** Adj

3 Context Elements for Adjectives

- Head noun = **Adj** N
- Adverb = Adv **Adj**
- Attached prep phrase(s) = **Adj** PP… (predicate adjective)

4 Context Elements for Adverbs

- Prep phrase (?) = **Adv** PP

5 Context Elements for Prepositions

- Noun object of prep = **Prep** N
- Prep phrase = **Prep** N PP

Tests on Context Elements

As noted, tests on the context elements specified in a transfer condition refer to feature values either hard-coded in the lexicon or assigned during analysis. In general,

- A test for part-of-speech value is the only obligatory test
- Boolean combinations of tests are permitted to the extent that the relevant MT systems support them.

Some tests on context elements are independent of part-of-speech designation, others are specific to nouns, verbs, etc. The following is an initial suggested list of features to be tested:

- Part-of-speech
- Canonical form of element (also as head of noun compound)
- Semantic type
- Syntactic type
- Natural gender
- Case/role
- Number
- Degree

Tests on the Source Word/Phrase

Feature values associated with the source word or phrase can serve as well as tests for transfer. Several refer to the broad text context of the source word/phrase, eg., value for subject area. Others, like tests on context elements, are either explicitly coded in the lexicon or assigned by analysis. Again, some of these features refer to all parts-of-speech, some are specific to part-of-speech.

- Semantic type
- Subject area
- Product
- Company
- Number
- Voice
- Case
- Tense
- Degree

Full Idiomatic Phrases

The user should be able to enter full phrases as a context element for the source word/phrase; this implies that the transfer condition is satisfied if the input source string matches word-for-word with the condition as it is stated, e.g., *trip the light fantastic, be in hot water.*

Heads of Compounds

Transfer conditions that are formulated for a specified source noun should be valid for compound nouns that contain the source noun as head.

Usage of Synonyms

A common wish-list element for MT users is the ability to specify synonyms as part of transfer conditions, e.g., *X is translated as Y in the context of Z or any synonyms of Z.* Since links based on synonymy are part of the Otelo DB specifications, using them to fill out transfer conditions is something that could be discussed further.

Target Transformations

In addition to stating conditions for transfer, the user should also be able to indicate cases in which the standard system handling of a particular string will not work given the context. In these cases

(at least some of them!), the user can define special transformations source-to-target that apply under the conditions specifically indicated by the user.

General Options

Users should have the option of assigning transfer to any element in the transfer condition statement. If the transfers that are assigned are not already in the lexicon, the user can be queried on whatever associated grammatical information is necessary to generate the correct form(s) for the transfer, e.g., *gender, morphological pattern codes, adjective position.*

List of Transformations

Transformations should be possible if the source word/phrase is one of the following parts-of-speech:

- Noun
- Verb
- Adjective
- Preposition

Noun Transformations

- Add preposition to context noun = **N** N -> **N** Prep N
- Delete preposition from attached PP; assign case/role to N = **N** Prep N -> **N** N
- Add determiner to N = **N** -> Det **N**
  N N -> **N** Det N
  N Prep N -> **N** Prep Det N
- Delete determiner from N = Det **N** -> **N**
  N Det N -> **N**N
  N Prep Det N -> **N** Prep N
- Add descriptive adjective = **N** -> Adj **N**
- Delete descriptive adjective = Adj **N** -> **N**

Verb Transformations

- Add noun argument; = **V** -> **V** N
  assign case/role to N
- Delete noun argument = **V** N -> **V**
- Add preposition to object N = **V** N -> **V** Prep N
- Delete preposition from attached PP; = **V** Prep N -> **V** N
  assign case/role to N
- Reorder cases/roles of argument N's = **V** N1 N2 -> **V** N2 N1
- Change voice of verb; = **V**(active) -> **V**(passive)
  adjust cases/roles of noun arguments **V**(passive) -> **V**(active)
- Add adverb = **V** -> **V** Adv
- Delete adverb = **V** Adv -> **V**
- Add predicate adjective = **V** -> **V** Adj
- Delete predicate adjective = **V** Adj -> **V**

2.1.1 Adjective Transformations

- Add adverb = **Adj** -> Adv **Adj**
- Delete adverb = Adv **Adj** -> **Adj**

2.2.4 Preposition Transformations

- Add determiner for noun object = **Prep** N -> **Prep** Det N
- Delete determiner for noun object = **Prep** Det N -> **Prep** N
- Add descriptive adjective = **Prep** N -> **Prep** Adj N
- Delete descriptive adjective = **Prep** Adj N -> **Prep** N

Note: As with the statement of transfer conditions, transformation statements should be relegated to administrators.

3.3.2.2 Synoptic table of the information types in the METAL lexicons.

Table 14: Lexical Information in the METAL lexicons.

	Entry component		Information content	Present
1	Headword		lexical form(s) of the headword: how the headword is spelt	✓
2	Phonetic transcription		how the headword (or variant form etc.) is pronounced (in *International Phonetic Alphabet*)	
3	Variant form		alternative spelling of headword or slight variation in the form of this word	✓
4	Inflected form		other grammatical forms of the lemma (headword)	✓
5	Cross-reference		indication of another headword whose entry holds relevant information, or some other part of the dictionary where this may be found	
6	**Morphosyntactic information**			
	a	Part-of-speech marker	part of speech of the headword (or the secondary headword)	✓
	b	Inflectional class	Inflectional paradigm of the entry	✓
	c	Derivation	Cross-part-of-speech-information, morphologically derived forms	
	d	Gender	Information about the gender of the entry in SL and TL	✓
	e	Number	Information about the grammatical number of the entry in SL and TL	✓
	f	Mass vs. Count	Information whether a noun is mass or count, in SL and TL	✓
	g	Gradation	For adverbs and adjectives	✓
7	Subdivision counter		indicates the start of new section or subsection (*'sense'*)	
8	Entry subdivision		separate section or subsection in entry (often called *dictionary sense*)	

9	Sense indicator		synonym or paraphrase of headword in this sense, or other brief sense clue indicating specific sense of SL or TL item	
10	Linguistic label		the style, register, domain, regional variety, etc. of the SL or TL item	✓
11	**Syntactic Information**			
	a	Subcategorization frame	(i.) Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.	✓
	b	Obligatority of complements	Information whether a certain complement is obligatory or not	✓
	c	Auxiliary	Which type of auxiliary is selected by a given predicate (in certain languages auxiliary selection is related to issues like unaccusativity, which on turn lies at the interface between lexicon and syntax)	✓
	d	Light or support verb construction	Constructions with light verbs	
	e	Periphrastic constructions	Constructions containing periphrasis, usage, semantic value, etc.	
	f	Phrasal verbs	Particular representation of phrasal constructions	
	g	Collocator	(i.) typical subject /object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented) etc.	
	h	Alternations	Syntactic alternations an entry can enter into	
12	**Semantic Information**			
	a	Semantic type	Reference to an ontology of types which are used to classify word senses	✓
	b	Argument structure	Argument frames, plus semantic information identifying the type of the arguments, selectional constraints, etc.	

	c	Semantic relations	Different types of relations (e.g. synonymy, antonymy, meronymy, hyperonymy, Qualia Roles, etc.) between word senses, etc.	
	d	Regular polysemy	Representation of regular polysemous alternations	
	e	Domain	Information concerning the terminological domain to which a given sense belongs	✓
	f	Decomposition	Representation of relevant meaning component, e.g. causativity, agentivity, motion, etc.	
13	Translation		TL equivalent of SL item	✓
14	Gloss		TL explanation of meaning of an SL item which has no direct equivalent in the TL	
15	Near-equivalent		TL item corresponding to an SL item which has no direct equivalent in the TL	
16	Example phrase (straightforward)		a phrase or sentence illustrating the non-idiomatic use of the headword, in a context where the TL equivalent is virtually a word-to-word translation	
17	Example phrase (problematic)		a phrase or sentence illustrating a non-idiomatic use of headword in a context where a specific TL equivalent is required (*i.e. an SL example which is easily understandable for the TL speaker, but presents translation problems for the SL speaker*)	
18	Multiword unit		(idiomatic) multiword expression (MWE) containing the headword *(the term MWE covers idioms, fixed & semi-fixed collocations, compounds etc*.)	✓
19	Subheadword *also* secondary headword		lemma morphologically related to the headword, figuring as head of a sub-entry (*subheadwords can be compounds, phrasal verbs, etc.)*	✓
20	Usage note		how the headword is used; 'macro' information which cannot appear at every appropriate entry; warning of cultural differences between the two languages; etc.	
21	Frequency		Information about the frequency of the entry	

3.3.3 Dictionaries of the Japan Electronic Dictionary Research Institute

3.3.3.1 Introduction

The Japan Electronic Dictionary Research Institute Ltd (EDR)(http://www.iijnet.or.jp/edr/) was established in April 1986, with an overall budget of 14 billion Yen covering the period up to the end of the fiscal year 1994.

EDR is supported by:

- The Japan Key Technology Center
- Fujitsu Ltd
- NEC Corporation
- Hitachi, Ltd
- Sharp Corporation
- Toshiba Corporation
- Oki Electric Industry Co Ltd
- Mitsubishi Electric Corporation
- Matsushita Electric Industrial Co Ltd

In addition to the lexical resources themselves, EDR also works on designing corpus building and processing tools, and on tools for creating and manipulating lexical data bases and knowledge bases. EDR is interested as much in tools for the lexicographer as in tools for the end-user (to customise or select dictionary material for use in a NLP system). The EDR corpora comprised some 20 Million sentences in Japanese and English.

The EDR English dictionaries were built with very little aid from native English informants, although efforts have been made to rectify this.

It is noticeable that the EDR dictionaries have been designed, implemented and constructed largely by computer scientists and engineers. There is no linguistic theory underlying the EDR dictionaries. This raises the serious doubt as to whether the information will be at all re-usable in a meaningful sense by theory-based NLP systems.

English descriptions are predicated largely on the needs of algorithms commonly used to process Japanese. They are also predicated on the types of descriptions traditionally used for Japanese. This leads to a symmetrical structure over the EDR dictionaries which is useful from the point of view of ease of maintenance and processing, however it has the undesirable effect, taken together with the lack of theoretical linguistic foundations, of leading to a blurring of boundaries between linguistic levels. This is seen particularly in the areas of orthography, morphographemics, morphosyntax and syntax. It is consequently very difficult to see how to relate the needs of a typical Western NLP

system that relies on the identification of various well-known linguistic levels to the data and their classification and description in the EDR dictionaries.

There is, it must be said, a lot of probably very useful surface observation of the cooccurrence of lexical elements in the English Word and Cooccurrence Dictionaries, derived from corpus processing, which should prove re-usable in the sense of being useful input for processes that may yield more theoretically adequate material.

It is difficult to judge the usefulness of the Japanese dictionaries, however it is assumed that these dictionaries have been built through consultation with NLP systems designers in the EDR investing companies, who presumably have some hope of re-using the dictionary information.

3.3.3.2 Overall Structure of the EDR lexical resource

There are several EDR Dictionaries. We deal here only with those containing general language: Japanese Word Dictionary (260,000 words), English Word Dictionary (190,000 words), Japanese Coocurrence Dictionary (900,000 phrases), English Cooccurrence Dictionary (460,000 phrases), Japanese-English Bilingual Dictionary (230,000 words), English-Japanese Bilingual Dictionary (160,000 words), Concept Dictionary (400,000 concepts)

EDR derived the dictionaries from 2 corpora of some 20 million sentences each, in Japanese and English.

The dictionaries are inter-related structurally in a complex fashion There is however a measure of redundancy in some dictionaries, as various parts are (conceptually) repeated from other dictionaries. For example, the bilingual dictionaries include surface-oriented information from the two word dictionaries.

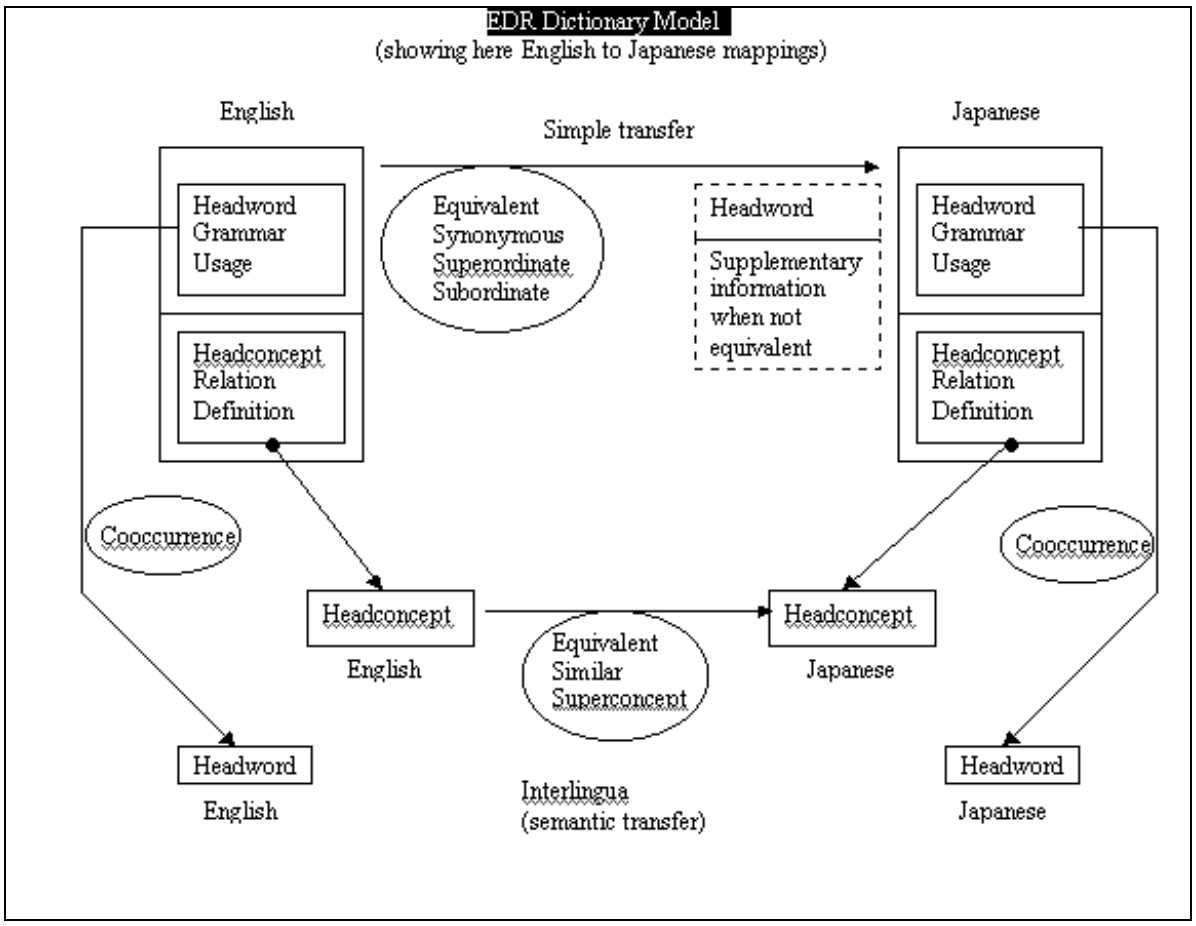


Fig. 14: EDR Dictionary Model

3.3.3.3 Name of Resource: EDR Japanese Word Dictionary

Organization and Structure of resource:

Structure of dictionary entry:

Headword information

- text form of headword (‘normal notation’)
- canonical form (‘retrieval entry’ – *the invariable portion of a string of characters – not equivalent to the word stem*)
- constituent information – *indicates where other words or phrases can be inserted in a compound headword.* Generalisation is achieved through use of **word classes** (drawn from the EDR Cooccurrence Dictionary) in the constituent expressions. By convention, ‘/’ separates the units of a compound word, ‘//’ indicates where a word may optionally be inserted, or where constituent order may change and ‘*’ stands for any word class.
- left and right side adjacency attributes – *indicate the possibility for joining morphemes* (a mixture of morphosyntactic, morphographemic and cooccurrence information). These attributes are for use in rules for both analysis and generation. Adjacency attributes can appear on both headword **and** the components of constituent information. The division into **left** and **right** reflects EDR’s *bidirectional connection method* which describes connectivity of a morpheme to its left and to its right.
- extra notation – For Japanese, gives the uninflected part of a headword in katakana when the pronunciation and normal notation are at odds (used for kana-kanji conversion and for determining word readings in text). For English, contains the entry word string with syllable markers to be used for hyphenation (not given however for compounds).

Syntactic information

- part of speech (includes phrasal categories for compounds)
- syntactic tree – *represents a structure of a compound word with its constituent words*. The tree can represent:

- optional and obligatory elements
- empty nodes where ill-defined modifiers may be inserted
- the boundaries of a constituent which can be moved to another position within the tree
- left and right adjacency attributes for the words of the compound

Word form information (any form not covered by the following will appear as a headword in its own right)

Japanese

word forms of conjugated words
conjugation type for –
  verbs
  adjectives
  adjectival nouns
  auxiliary verbs
compound words
conjugation constraints

English

inflection information for verbs, nouns, adjectives & adverbs
case and number information
special inflection information (for irregular forms)
words modifiable by determiners and adverbs
modifiers of nouns and adjectives
information on syntactic dependency
special treatment of nouns in number agreement

Surface case of predicates

Aspect

Categorisation of verbs

information on function words (particles, particle- equivalents, formal nouns, auxiliary verbs, etc.)

Usage information

Frequency of occurrence of the headword in the EDR corpus

Pronunciation

Japanese

pronunciation in katakana
stress marked by symbols
no distinction between voiced and nasal
no special treatment of ‘double consonants’ or long vowels
standard Tokyo accent
rudimentary inflection for compounds and idioms

English

pronunciation in IPA
optional sounds are bracketted
accents marked by diacritics
syllable division indicated

Semantic information

headword and definition of a single word entry OR headwords of a compound entry and their respective definition, plus labelled relational structure of the compound

Ordering of senses: one entry refers to one sense

3.3.3.3.1 Comments on EDR Word Dictionary

The EDR Word Dictionary records largely surface information on wordforms. The semantic field of the dictionary entry contains a minimum of information: effectively a definition of the concept referred to by the entry headword. This field is used to index into/from the EDR Concept Dictionary, where fuller semantic information may be retrieved, and where interlingual translation may be effected. The headword information field (a complex field) of the Word Dictionary also allows indexing into/from the EDR Cooccurrence Dictionary and into the EDR Bilingual Dictionary.

The dictionary stores fullform words, as they occur in text. Within an entry, a canonical form is stored, however this does not necessarily represent what a linguist would recognise as a stem, but is simply the invariant part of a character string, common to the several variants or realisations of a word. It is possible that the indexing and hence organisation of the entry is in fact different to that described in the available report (e.g. several text wordforms may map to an entry with one canonical form, with its associated information).

The needs of Japanese for kana-kanji conversion are accommodated in a special field, the ‘extra notation’ field, whose contents are also used to aid in disambiguation of senses.

There is an extensive amount of detail on ‘adjacency’ information. That is, for each lexical unit, information is given on possible elements to the left and right of the lexical unit. Such information on context is stored to enable the writing of morphological rules. The type of contextual information stored varies over many different linguistic levels. E.g. morphographemics, morphosyntax,

punctuation, syntax. Each type of element that could occur to the left or right is given a unique category code.

As for derivational morphology, this is largely missing from this dictionary as far as can be told (all relevant linguistic labels are given *in extenso* with examples in the documentation, which permits us to deduce the lack of derivational morphology information). There is a minimal treatment only: E.g., for English, there is a label for prefix (ECF1) available in the right side adjacency attribute set; and a label for prefix (EPF) in the syntactic information regarding parts of speech; there is also a syntactic part of speech suffix label (EUN) which is however apparently restricted to the coding of lexical units denoting units of measure such as 'cm' and 'kg' – in other words, 'cm' is coded as a suffix.

The description of English is heavily influenced by that for Japanese held in the Japanese Word Dictionary. There is for example a great detail on surface context of words (but compare also the information of the Cooccurrence Dictionaries). The reason given is that the description is done in this way to enable a NLP system capable of processing Japanese to re-use the same algorithms and techniques for English. Therefore the description of English effectively assumes that there is no word boundary information available for example in the sentence string being analysed. There is no indication that the description is based on a theory of linguistics.

Syntactic information gives among other elements part of speech. A phrasal approach is adopted to the encoding of compounds, which for several years now has been rejected as inadequate by mainstream linguistics. Compounds receive a separate tree structure, which indicates possibilities of optionality of arguments, insertion of modifiers, etc. Although all the available examples of compounds were of complex expressions such as phrasal verbs, phenomena such as noun-noun compounds "N+N...+N" are also catered for. Other information included in the syntactic description concerns conjugation information, surface case of predicates, aspectual information for verbs, information on function words, usage and pronunciation.

It should be noted that, in Japanese linguistics, there is a fuzzy distinction between morphology and syntax, due to the nature of the writing system. Therefore, the distribution of what western linguistics would recognise as morphological and syntactic information over the lexical entry appears odd, whereas to a Japanese linguist this is perfectly natural. Nevertheless, it is true to say that the linguistic description appears to be couched in terms of 'naive (traditional) linguistics', and does not therefore make appeal to any theoretically based notions.

Minimal semantic information is included to enable the identification of a sense by a human. At the computational level, the word dictionary entry contains a mapping to the concept dictionary where the bulk of semantic information is stored.

The EDR dictionaries are designed to be re-usable, however it is quite unclear to what extent the Word Dictionary would be re-usable in a theory-based NLP system, e.g. a NLP system based on JPSG (roughly: the Japanese equivalent of GPSG), or, more generally, any NLP system which implemented a standard Western view of processing character strings which is at odds with the EDR assumed view. It is likely that much of the contextual information could be extracted and re-expressed as general rules. The division and distribution of a particular type of information (e.g. morphological) over several EDR-specific 'linguistic' levels is a barrier to re-usability that would have to be overcome.

3.3.3.4 EDR Japanese and English Cooccurrence Dictionaries

Organization and Structure of resource:

Structure of dictionary entry:

Headword-1 information – Identical to that contained in the corresponding Word Dictionary. Cooccurrence relation between Headword-1 and Headword-2 –

The syntactic role of two words/morphemes is expressed by a cooccurrence relation. There are as many separate coccurrence dictionary entries as possible cooccurrence relations between any given pair of Headwords. However, a cooccurrence relation can also describe a relation between *groups* of words/morphemes. Words/morphemes therefore can be grouped into classes for the purpose of establishing cooccurrence relations.

Note: *Extra notation* for a Headword is not given in the Cooccurrence Dictionary – this however is available in the associated Word Dictionary.

Ordering of senses: one entry refers to one cooccurrence relation between 2 Headwords (or *classes* of Headword).

3.3.3.4.1 Comments

There is no available publication devoted entirely to the Cooccurrence Dictionaries (as there is for the other EDR dictionaries). This leads one to surmise that either there is little more to be said than what appears in overview publications, or there has been little work in fact done on the Cooccurrence Dictionaries. As little is said about progress on the Cooccurrence Dictionaries, this reinforces the latter interpretation. However, Nakao (1990), while discussing techniques of extracting data from the EDR corpus, notes that information for the Cooccurrence Dictionary is obtained automatically to a large degree.

EDR defines cooccurrence as follows:

> *When a specific element, such as a morpheme or phoneme, co-occurs with another element of the same type in one word, phrase or sentence without grammatical deviation, these two elements have a cooccurrence relation.*
>
> (EDR, 1990a:17)

The use of the phrase "of the same type" renders this definition somewhat obscure, as does the usage of 'grammatical'. One might well prefer 'pragmatic' to 'grammatical', as does EDR elsewhere when it is noted that the Cooccurrence Dictionary gives

pragmatic information for generating a sentence with natural wording

(EDR, 1990a:3)

The stated role of the Coocurrence Dictionary is to aid in the selection of translation equivalents. Where there are several possible surface realisations of a concept dependent on context (i.e. cooccurrence possibilities), then the Cooccurrence Dictionary allows the correct choice to be made. Thus, if a concept has been previously identified such as DRIVE and there are several possible surface realisations, then the Cooccurrence Dictionary is accessed to resolve the ambiguity. To take an English example here, we may find headwords corresponding to DRIVE such as 'drive', 'ride', etc. The (English) Cooccurrence Dictionary would then reveal cooccurrence possibilities. For example, we may find the cooccurrence entries (drive,@objective,car) and (ride,@objective,bicycle). The concept DRIVE will, we assume here, stand in an objective relation to another concept, say, AUTOMOBILE. Matching of this conceptual structure against the headword cooccurrence possibilities of the Cooccurrence Dictionary will allow the headword 'drive' to be chosen in this instance, as opposed to 'ride', in other words, this allows generation of "X drives a car" as opposed to the non-preferred "X rides a car".

A fuller translation based example follows below.

Assume the following interlingual concept relation representations (see section on the EDR Concept Dictionary):

<catch> – object <cold> (to catch a cold)

<catch> – object → <flu> (to catch flu)

In Japanese, the concept <catch> is expressed by different words depending on the object concept. The Coocurrence Dictionary provides the following information:

(kaze, @objective, hiku) where 'kaze' = 'cold' and 'hiku' = 'catch'

(ryukan, @objective, kakaru) where 'ryukan' = 'flu' and 'kakaru' = 'catch'.

The above surface cooccurrence information allows selection of appropriate translation equivalents in Japanese, yielding:

"catch a cold" → "kaze wo hiku"

"catch flu" → "rkukan nu kakaru".

The relationship between cooccurrence information, adjacency information and syntactic information is unclear, given the lack of theoretical basis in the EDR Dictionaries. This is especially true for the English Dictionaries. Reference to the English Word and Cooccurrence Dictionaries shows a lack of real distinction between what a linguist would recognize as morphological, cooccurrence and syntactic information: that is, we find, e.g. in the English Cooccurrence Dictionary the fact that 'un-' can cooccur with 'fortunately', and 'an' can cooccur with 'umbrella'. This information is explicitly recorded via a cooccurrence relation label, despite

the fact that in the former case we are dealing with a phenomenon from derivational morphology and in the latter with a syntactic phenomenon of determination. Such information is presumably also expressed in different form in the Word Dictionary (various notational mechanisms are available for this) in terms of e.g. adjacency attributes. Given the structure of the Cooccurrence Dictionary, adjacency attributes are present in an entry, being part of the information recorded for each Headword pair in an entry. Derivational morphology however does appear to be dealt with mainly in the Cooccurrence Dictionary proper as opposed to the Word Dictionary (i.e. by relation labels between Headwords), although there are labels in the Word Dictionary for recording of affixal information. Derivational morphology is apparently restricted to simple statements of adjacency in the Cooccurrence Dictionary.

In general, there appears to be a possibility of a certain (even large) amount of redundancy between Cooccurrence Dictionary information (expressed through relation labels) and Word Dictionary information (expressed through several means, e.g. adjacency attributes).

In conclusion, we note that the bulk of data for the Cooccurrence Dictionaries appear to be derived automatically from the EDR corpus, with some human intervention to tidy up manifestly wrong or quite useless (too general) cooccurrences. We further note that (English) compound words, such as noun-noun compounds, appear to be handled exclusively in the Cooccurrence Dictionary (insofar as their surface characteristics are concerned), although this could be simply the effect of choice of example in the relevant documentation.

As regards re-usability, the available very limited description does not allow any accurate assessment to be made. In particular, one would require details of the cooccurrence extraction algorithm, plus exhaustive information on cooccurrence relation labels, before being able to form a judgement as to re-usability. It is however likely that many relations have been established, which could prove if not directly re-usable (given the lack of theoretical basis prevalent in the EDR dictionaries) at least indirectly re-usable after manipulation.

3.3.3.5 EDR Bilingual Dictionaries (Japanese-English and English-Japanese)

Organization and Structure of resource:

Structure of dictionary entry:

Source language headword information (identical to that in the corresponding Word Dictionary)

- Inter-lingual correspondence label – this field contains the label which gives the bilingual (unidirectional) correspondence between a pair of headwords (English–Japanese or Japanese-English). There are four values available, namely:

equivalent relation

synonymous relation

superset relation

subset relation

Target language headword information – same type of information as for source language entry, plus 'supplementary explanation' for non-equivalent headwords

Notes on the correspondence relations:

The relations between corresponding headwords are described in an ordered fashion.

That is, preference is given to describing equivalence relations. If no equivalence can be established, then a synonymous relation is specified. Failing synonymy, a superset relation is sought, and failing that a subset relation is established.

- equivalence relation: indicates there is a "nearly one-to-one correspondence [...] In many cases, a [source] word can be replaced by a corresponding [target] headword".

- synonymous relation: here the source headword "differs enough from its corresponding [target] headword that it cannot be regarded as an equivalent relation". Mistranslation would result if a target headword were used for a source headword under the synonymous relation without supplementary explanation.

- subset relation: indicates that the source headword "covers a wider range of concepts than the corresponding [target] word". Thus, target headwords linked to source

headwords by this relation can be used “only in specific situations in which the [source] headword is used”.

- superset relation: indicates that the target headword covers a wider range of concepts than the source word. “A corresponding [target] headword can be used only when what it represents is limited”.

Notes on the target headword information (corresponding headword):

This contains the same type of headword information as for the source language field, plus additional information on ‘supplementary explanations’.

EDR have set up several criteria to guide selection of target headword for inclusion in the Bilingual Dictionary:

- a target headword with corresponding grammatical features is to be preferred. Note: this enables a client system to implement a *simple transfer* strategy.

- general-purpose headwords are to be preferred: this is to avoid too specific translation in specific contexts.

- target headwords that are ‘compact’ are to be preferred: this is to avoid the use of explanatory phrases and complex phrases, seen especially in a preference (in the English-Japanese Bilingual Dictionary) for target headwords that are Chinese compounds, rather than Japanese paraphrastic expressions for the same concept.

- if there is no possible target headword, due to lexical gaps, then the source language headword is borrowed for use as a newly-created target headword (with appropriate conversion to target language conventions).

- if no equivalence relation can be established, then ‘supplementary expalanations’ are added to gloss the type of relation (which then must be one of synonymy, superset or subset).

Notes on ‘supplementary explanations’:

A supplementary explanation (or explanations) is recorded in the target headword field of the Bilingual Dictionary in the case where no equivalence relation can be established. This explanation

supplements the relation label of superset or subset, and is given in a combination of a coded and textual form. There are three codes used:

1 – indicates a narrowing down of the meaning of the target headword

2 – indicates a restriction on the situation in which the target headword is used, or to indicate usage

3 – explains the meaning of the target headword

Code 1 is apparently used only to gloss a target headword in a superset relation; codes 2 and 3 are apparently used only to gloss a target headword in a subset relation.

It is unclear from the documentation what happens in the case of a synonymy relation.

The format of a supplementary explanation is:

(code: textual explanation in the target language)

Supplementary explanations are "described in natural expressions so that they can also be used as part of the output sentences".

Ordering of senses: one entry refers to one bilingual correspondence. If a source language word has several translations, then a new entry is set up for each.

3.3.3.5.1 Comments

It is important to note that the EDR Bilingual Dictionaries establish bilingual correspondences between **words**, not concepts. They are clearly intended to support bilingual NLP applications that exploit the notion of *simple transfer*, that is, where word-for-word translation is practised (based on a compositional analysis, typically, as in Eurotra), and where the target expression is constrained to be of the same grammatical category as the source expression.

Nevertheless, as soon as description departs from the realm of equivalence relations (effectively equivalence of two words in context), correspondence between concepts necessarily enters into consideration. The EDR Bilingual Dictionaries note unidirectional relations of equivalence, synonymy, superset and subset between source and target **words**, although it is apparently the case that these relations are set up on the basis of conceptual criteria. However, the emphasis is still on words. The superset relation for example indicates that a target **word** covers a greater range of concepts than the source word *viewed from the point of view of the source language* — the target

language may not in fact *recognize* more than one concept, but from the point of view of the source language we note that the target language word can be used to refer to more than one source language concept. If therefore a source language word is seen to be polysemous or homonymous with respect to the target language, although it may not be seen to be such in the source language, a new bilingual entry is constructed for each correspondence.

An unfortunate aspect of the EDR documentation is that it obscures the role of certain of the interlingual relations due to badly chosen examples. This can be seen in the description of e.g. the subset relation where it is claimed Japanese 'mugi' maps to three different (narrower) English words ('barley', 'wheat' and 'rye'), there being no equivalent English generic word. The associated discussion however notes that "mugi refers to grain". Presumably, the argument is that 'grain' in English has a wider reference than 'barley', 'wheat' and 'rye', unlike Japanese 'mugi' which refers only to these three cereals. This is however not stated explicitly. The point however is well-taken, namely that there is a need to record interlingual subset mappings and moreover to gloss the meaning of the target via 'supplementary explanations'.

More problematic is the exact nature of the synonymy relation. The documentation states that all non-equivalent mappings are further glossed via 'supplementary explanations', and yet enters into detail only on 'supplementary explanations' for subset and superset relations. Indeed, the codes 1 ...3 are distributed only over these two relations (see above). This leaves the synonymy relation quite underdescribed. EDR itself appears to be undecided as to the usefulness of the synonymy relation, as we are informed that the number of words that have translation words with the synonymy relation is only about 300 of 400,000 words in the Word Dictionaries. EDR further informed us that the relation might be changed into some other relation, however current documentation still mentions the synonymy relation.

The technique of using 'supplementary explanations' complements the use of all non-equivalence relations[5], and is meant to indicate to the target language user how the target expression is constrained, by offering a coded indication of modification, together with a brief textual explanation. The textual explanation is constructed in such a manner that it can be directly incorporated as part of the output text of some system (all things being equal) – however it is not intended that such supplementary information should always be output. We note a somewhat unconvincing example, namely the correspondence (superset relation here):

Japanese	English
keshigomu	(1:pencil) eraser
kokubankeshi	(1:blackboard) eraser
inkukeshi	(1:ink) eraser

It is the case that English can quite happily admit 'pencil eraser' as a compound word, as it can also 'blackboard eraser' and 'ink eraser'. The problem is then one of dealing with reduced forms of

[5] but appears to be in fact used only for subset and superset relations – which may just be due to a lack of attention in documentation.

compounds (or, in a different world, identifying concepts and their names as realised in different environments). In other words, it is the case that ‘keshigomu’ can map to 2 English strings, namely ‘eraser’ and ‘pencil eraser’, which are one might say textual variants. The examples given however omit to mention the possibility of a mapping to a non-reduced compound. This does not mean that this mapping may not exist somewhere (presumably as an equivalent relation mapping). If such a mapping to a non-reduced compound does not exist, however, then there is a strong likelihood that odd translations will be produced whenever the non-reduced compound appears in a text. The result typically produced for a translation of an English sentence containing the string “pencil eraser” into Japanese would then be something like (glossed in literal English to aid clarity):

“...enpitsu-yoo-no (3:enpitsu) keshigomu ...” “...pencil (3:pencil) eraser ...”.

Here, we assume the mapping of the example entry ‘eraser’ → ‘keshigomu’ and also the mapping for an entry ‘pencil’ → ‘enpitsu’. The elements ‘yoo’ and ‘no’ are particles.

We hasten to emphasize this is a rather naively constructed example (Japanese generation may prefer oher methods for expressing the modification relation between the lexemes ‘keshigomu’ and ‘enpitsu’). However, it is quite likely that redundant information will be generated in the translation. Naturally this depends heavily on many other details. We wish here simply to state a possible shortcoming of the bilingual mapping strategies adopted in the EDR bilingual dictionary. This is a problem for other dictionaries as well as the EDR ones. It is also the case that one may argue that a dictionary should provide information that is somewhat redundant, and let NLP system strategies filter out the actually required information. What can be noted in the EDR dictionary however is that in the case of superset and subset relations, these are not totally formally described: there are ‘supplementary explanations’ that are given in free text form, and that, it is stated, can be output as part of the target translation.

There is clearly an interaction to be considered between supplementary explanations and cooccurrence information. It may be the case, for example, that the target language cooccurrence dictionary could help select the correct translation by matching the wider sentential context against coocurrence dictionary entries. This interaction is not discussed in EDR reports.

3.3.3.6 Name of Resource: EDR Concept Dictionary

Organization and Structure of resource:

Overall, the dictionary has the form of what EDR call a ‘hyper-semantic network’ (a semantic network which contains semantic networks embedded in its nodes, and in which nodes in embedded networks are allowed to form links outside their embedded network).

The detailed structure is described and commented on below.

3.3.3.6.1 Comments

The Concept Dictionary is viewed as the ‘key dictionary’ by EDR. It is intended to provide interlingual conceptual information suitable for at least Japanese and English. The methodology chosen to develop the dictionary (see below) is intended to ensure that the concept dictionary can also be used for other languages, as it is intended to reflect ‘universal knowledge’ that is independent of language.

The main objective of the concept dictionary is to provide a means of translating between English and Japanese, in situations where syntactic or cooccurrence methods fail, or where semantic information must be recovered or inferred which is perhaps missing in the surface sentence (cf. the case of elided post-prepositions in Japanese, or word order changes in English).

Consultation of the concept dictionary allows “precise recognition of the semantic relationships between words” to be achieved. In a machine translation environment, use of the concept dictionary “extends the range of appropriate wording and enhances the variety of expressions in the target language”. The concept dictionary comes into play when there is a need for the equivalent of complex structural transfer, as well as when higher level semantic information must be accessed. The comparison just made reflects a particular strategy, which need not in fact be followed. The EDR dictionaries are declarative knowledge sources which can be used individually in different ways by different strategies. For example, a particular NLP system might operate largely with conceptual information after accessing the appropriate concepts, and pay little attention to e.g. cooccurrence information.

There are two main parts to the Concept Dictionary, namely the *Concept Description* and the *Concept Classification*.

The *Concept Description* is a horizontal description. It is based on analysis of sentences from the EDR corpus, and yields a network of conceptual case relations among concepts (agent, object, implement, location, etc.), with supplementary relations to express attributes such as aspect and set values (such as ‘generic‘, ‘all’, etc.) (see below). These relations are referred to as ‘conceptual relations’. Note that as we are dealing with concepts, the ‘case’ roles are not to be understood as giving e.g. noun arguments of verbs, but giving instead relations between e.g. events and objects. There is no indication given whether EDR has developed its own set of cases, or borrowed these from elsewhere.

The *Concept Classification* is a vertical description, resulting from top-down human analysis, reflecting an organisation via IS-A links (here: 'kind-of'). Other 'semantic relations' such as 'part-of', 'equivalent' and 'similar' are also used, however the primary organisation is done in terms of 'kind-of' links. This structure allows reduction of information – it does away with redundant information that can be inferred or inherited.

The Concept Classification and Concept Description are different *views* over the Concept Dictionary, differentiated mainly by the type of relation involved: 'conceptual relations' link items of the Description, whereas 'semantic relations' link items of the Classification. However, the overall structures of these two views are the same. In this sense, the structural organisation of the Concept Dictionary is kept relatively simple.

The central unit of the Concept Dictionary is the 'headconcept'. This is an identifier for an individual concept expressed by a word in the Word Dictionary. Word Dictionary entries contain a field which holds the headconcept a word is related to. Polysemy leads to separate headconcepts for each sense of a word.

A concept indicated by a headconcept is "an abstract essence of the common meaning of a word, free from shades of meaning generated under various situations". This means that concepts are selected and represented which do not rely on any viewpoint or intention of speakers, and do not rely on contextually or situationally dependent contexts. A concept in EDR's view is "a class of images consisting of common attributes and components regarded as independent of the context or situation". For example, the concept <chair> is taken to relate to: a set of images of types of chair a set of images created by the group of attributes which describe a chair

Headconcepts are identified on the basis of common sense. Separate concepts are recognised for derived, figurative and metaphorical meanings. It should be noted that concepts are not defined by reference to sets of primitives. EDR explicitly rejects this method, as it is considered to be not proven. EDR has similarly rejected an approach which links each word of a language to a concept in a one-to-one relation. This latter approach would not allow establishment of an interlingua, at least not directly. In essence, EDR views concepts as acquiring meaning through relation to other concepts. Definition of concepts and construction of relational structures must therefore proceed hand-in-hand. The objective is to describe as many concepts initially as possible, to describe their relations to each other, and subsequently to modify the description of concepts in the light of possible relations. This leads to a reductive approach (as does the primitive-based method) which is however more likely to be effective, in EDR's view, as a set of useful and well-defined *interlingual* concepts will arise out of massive analysis of data, massive specification of relations, and massive re-appraisal of initially-proposed concept descriptions (indeed, re-appraisal would probably take place on several occasions in a cyclic methodology).

The methodology of concept and headconcept selection and identification is important. The objective is to arrive at a set of interlingual concepts defined as the union of a set of language-independent concepts and a set of language-dependent concepts. Miike (1990) enters into detail on this process. The *definition* is central to the methodology. For the purposes of the following description, one may think of a headconcept simply as a definition (it is actually a definition – or other phrase identifying the concept – *cum* identifier). In summary, the set of headconcepts is arrived at as follows: a headconcept is set up for each word in a language Word Dictionary, independently of other headconcepts (i.e. a definition is written). For every headword paired with one of its headconcepts, lexicographers attach other words that *subsume* the concept represented by that headconcept. *The words can come from more than one language*. Note that subsumption is typically the case as there will be often no direct one-to-one correspondence possible between concept and word. One therefore aims for the most specific concept that is more general than the

meaning of the word under consideration. A listing is automatically produced of all headconcepts of the words noted by lexicographers for each headword-headconcept pair treated in the above step. This yields a group of headconcepts and associated words with a supposed equivalence relation holding between them. Lexicographers select one headconcept from each group that represents a concept common to the group of headconcepts. This step is carried out *on the basis of considering headconcepts alone*, in isolation from their headwords. This is to avoid unification of concepts on the basis of word-influenced senses as far as possible, and to render the headconcepts as language-independent as possible.

If no appropriate common headconcept can be selected for a group, then either a new headconcept is created (i.e. a new definition is written covering the concept circumscribed partially by each existing definition (headconcept)), or the group is deleted (in the case of groups containing totally disparate headconcepts, as can happen). Once each surviving group has received a single common headconcept, groups with similar or identical common headconcepts are examined to determine whether they should be conflated or otherwise differentiated. The previously removed headwords are re-instated for each group, and lexicographers asked to match each headword in a group with the group's common headconcept. Some refinement or replacement of the headconcept may take place at this time, in the light of the information brought in by the headwords now being available. This methodology is claimed to yield a set of headconcepts that has been elaborated largely in isolation from headwords. This claim is not without validity, however it requires a rather detailed and well-tried set of guidelines, to help lexicographers work in a way which is quite foreign to them. Normally lexicographers proceed from word to concept, whereas in this instance they proceed from concept to word (in much the same way as terminologists work). This work, in the context of large-scale conceptual resources for NLP, is innovative, and in the current framework of EDR research there is a certain amount of faith being invested in its ultimate validity and usefulness. For example, development of the English-based set of headconcepts involved unification of headconcepts being undertaken while the development of guidelines for headword attribution and indeed the feasibility of such attribution had not been worked out. It should be noticed that the methodology has been described in summary fashion above: there are many steps in fact necessary to resolve certain types of case.

We have entered into some detail here as it is necessary to understand the methodology of headconcept unification in order to appreciate the core nature of the Concept Dictionary.

It is this methodology which results in a set of headconcepts which can be said to be the union of language-dependent and language-independent concepts.

The Concept Dictionary is logically organised as a set of 'conceptual relation representations' (CRR). A CRR is a 'hyper-semantic network' (a semantic network which contains semantic networks embedded in its nodes, and in which nodes in embedded networks are allowed to form links outside their embedded network). A CRR can represent a simple conceptual structure, based on the simplest type of relational entity (the 'concept entry') or it can represent complex structures, containing various combinations of concept entries and representations of compound concepts. It is a recursive structure.

The units that contribute to building complex CRRs typically have the form:

<concept name>[<internal structure>]

where <internal structure> is a potentially recursive structure consisting of combinations of concept entries, single headconcepts and units with further internal structure.

Single headconcepts are the 'leaves' of CRRs. As such, they are defined in terms of themselves, as follows:

<animal called bird>[<animal called bird>]

EDR does not believe in setting arbitrary limits to the downwards expansion of its ontology: this reflects the view that a useful set of headconcepts will result from cyclic refinement and adjustment – there is no notion of a 'demonstrably complete ontology'.

Formally, a CRR is described as:

<CRR> ::= <concept name>[<internal structure>] | <concept name>[<concept name>]

<internal structure> ::= <CRR>* <concept entry> <CRR>*

<concept entry> ::= <binary relation> | <unary relation>

binary relation> ::= <concept reference> – <attribute>{/ certainty factor>} → <concept reference>

<unary relation> ::= <concept reference> – <attribute>{/ certainty factor>} → *nil* | *nil* – < attribute>{/ certainty factor>} → <concept reference>

<concept reference> ::= <concept name>* | [<internal structure>]

<certainty factor> ::= 1 | 0

<concept name> ::= <sentences, phrases, words for identifying concept> | '<' <headconcept> '>'

<headconcept> ::= <identifier and definition for concepts descibed in word dictionary>

<relation label> ::= <identifier of relations between concepts>

<attribute> ::= <delimiter of concept range>

Note: this is only a *partial, idealised* grammar of a CRR, but correct enough to indicate the major structures involved. The complete grammar can be found in the EDR Technical Report TR-027 *Concept Dictionary*, on page 12.

A *concept entry*, which shows a conceptual or semantic relation between two entities of the ontology, has the following general shape:

concept_reference^1 – relation → concept_reference^2

A relation can be either a *conceptual relation* (agent, object, ...) or a *semantic relation* (kind-of, equivalent, ...).

The *basic* concept entry states a relationship between headconcepts:

headconcept^1 – relation → headconcept^2

headconcept^1 is said to be the 'centre of relation' and headconcept^2 the 'object of relation' (to be distinguished from 'object relation'). Centre of relation headconcepts refer to events (e.g.

movement, action, change) and properties (e.g. shape, weight, colour). Object of relation headconcepts refer to e.g. physical objects, abstract things, human and animate concepts – they are classified on the basis of their relation with centre of relation concepts.

An example of a basic concept entry is:

<eat> – agent → <bird>

<...> is used simply here to indicate a headconcept.

The above expresses the fact that "the concept bird is an agent of the concept eat" (in an undifferentiated sense here: we have left out other information such as aspect and level of genericity, for example).

Each relation has an associated *certainty factor*, which indicates whether the relation given by the concept relation label is either possible (factor value = 1) or not possible (factor value = 0). When the factor value = 1 it can be omitted (i.e. 1 is the default value). In the above, the relation 'agent' could have been given an explicit certainty factor of 1: *agent/1*.

The use of 0 is somewhat dubious, in our view, as we see that the indication of an impossible relation is essentially a highly strategic decision, that may not have well-founded criteria for use. An example given in the EDR literature is:

<eat> – agent/0 → <stone>

i.e. that "stones don't eat". There is no indication given as to when or how such 0 factors should be used. It would appear to be highly unlikely that all impossible relations are explicitly marked. This leads one to suppose 0 factors are used *strategically* to avoid potential clashes and ambiguities – which if they exist would presumably indicate some failure in adequate discrimination of concepts and/or the relations between them.

The notion of concept in the EDR dictionaries extends to that of a *compound concept*: in fact, CRRs will typically represent the compound concept represented by a phrase, or sentence. Headconcepts will have been initially gathered typically by consideration of individual concepts. Compound concepts are constructed on the basis of corpus analysis, as "actual sentences are the best means for judging the existence and types of concept relations". Relations between concepts are determined on the results of automatic corpus processing (morphological and syntactic analysis – which increasingly use the growing EDR dictionaries). Lexicographers are presented with subtrees showing various syntactic modification relationships over parts of sentences. Conceptual relationships are specified on the basis of these. If a particular relationship can be in fact inferred by appeal to the existing dictionary conceptual structure and rules of inference and inheritance, etc., then that particular relationship is not recorded. CRRs must be able to describe any possible concept (compound concept) and moreover similar concepts (compound concepts) should have identical CRRs.

Compound concepts will typically describe a variety of relationships among constituent concept entries or embedded compound concepts. A simple concept entry can form the basis of a compound concept. Thus a very basic *compound concept* would look in full like:

<a bird flies>[<to fly in space> – agent/1 → <an animal called bird>]

which describes the compound concept <a bird flies>.

Interestingly, the following causes no problem, even though the nature of flying and the nature of the agent are conceptually different:

<an aeroplane flies>[<to fly in space> – agent/1 → <a transport means called aeroplane>]

Here the same agent relation is used: any ambiguity will be resolved by appeal to the concept classification and to combinability possibilities of the component concepts. In other words, general relations can be used, disambiguation being effected by other means.

A further example is:

<an apple is red>[<red colour> – object → <fruit called apple>]

A more complex example is:

<sumo wrestlers drink much alcohol> [<to drink> – agent/1 → <wrestlers of Japanese wrestling>, <to drink> – object/1 → <alcohol>, <to drink> – quantity/1 → <a large volume>]

Here, we have shown only one level of embedding. Note that the main conceptual relationships have been extracted and made explicit (there may of course be others).

It is clearly noticeable that compound concepts of the last type are approaching full sentence representations. This is a point which is somewhat unclear in the EDR literature. We will return to this below.

Due to the existence of the concept classification (giving an IS-A network), a reference to a concept in a CRR can be regarded as a reference to an entire class of concepts: “when a concept appears in the CRR, it is regarded as representing one subclass of a class. In this case, the class itself is also regarded as one of the subclasses”.

Attributes of concepts are defined as super-classes in the concept classification. This implies that descriptions can be kept within reasonable bounds, otherwise for each different attribute, a new concept would have to be set up. This is standard practice in knowledge base design. For example, we could have a concept of <institute which is a building> but instead we find in the classification:

<institute> – kind-of → <building>

If we subsequently are asked to verify, in the course of processing a sentence in an actual NLP application: <build> – object → institute

we can verify this from the above kind-of link and also from:

<build> – kind-of → <construct> <construct> – object → <building>

As noted above, concept entries can describe relationships between single headconcepts. They can also describe relationships between various combinations of headconcept and complex concepts (represented by an embedded CRR):

headconcept^1 – relation → headconcept^2

headconcept – relation → embedded_CRR

embedded_CRR – relation → headconcept

embedded_CRR – relation → embedded_CRR

One therefore finds concept entries such as:

[<to drink much alcohol> – kind-of → <to carouse>]

where <to drink much alcohol> is an embedded CRR, represented by the compound concept:

<to drink much alcohol>[<to drink> – object → <alcohol>, <to drink> – quantity → <a large amount>]

Note here the *list* of concept entries associated with the compound concept name <to drink much alcohol>.

<to carouse> is simply:

<to carouse>[<to carouse>].

By exploiting embedded CRRs and the different types of relations, we can build up such representations as:

<a person borrows a thing from a person>[<to borrow> – agent → <person>1, <to borrow> object → <thing>, <to borrow> – source → <person>2]

<a person lends a thing to a person>[<to lend> – agent → <person>1, <to lend> – object → <thing>, <to borrow> – source → <person>2]

[<a person borrows a thing from a person> – equivalent → <a person lends a thing to a person>,

<a person borrows a thing from a person> <person>1 – equivalent → <a person lends a thing to a person> <person>2,

<a person borrows a thing from a person> <thing> – equivalent → <a person lends a thing to a person> <thing>,

<a person borrows a thing from a person> <person>2 – equivalent → <a person lends a thing to a person> <person>1]

EDR employs various notational devices in order to simplify representations. We do not address this issue here. The EDR formalism allows further for indication of scope of reference within CRRs, according to explicit rules. There are various alternative conventions available to express scope.

Again in order to simplify descriptions, EDR employs a small number of 'pseudo-relations' such as 'possessor' which replace frequently occurring sets of relations. Thus, the complex:

<taro's book>[<possess> – object → <book>, <possess – agent → <taro>]

can be alternatively encoded as:

<taro's book>[<book> – possessor → taro]

Regarding *unary relations*, these are used to indicate attributes of concepts. Such attributes are divided into 'aspect attributes', which are drawn from the set {begin, progress, end, continue, state} and 'set attributes', which are drawn from the set {generic, specific, some, all, not}.

Examples are:

<to be walking>[<walk> – progress → *nil*]

<apple> – specific → *nil*

<apple> – generic → *nil*

The latter two reflect the difference between a specific instance of an apple (as in "I like this apple") and a generic notion of apple (as in "I like apples").

In the concept classification, object concepts can be specified for certain *attributes*.

Such attributes are set up based on the type of relation label that can link them to particular types of 'centre of relation' concept.

Thus, on the basis of the link between, say, <person> or <animal>, a relation label agent and a < controllable action> concept, one may set up <person(human)> or <animal(animate)>. This particular area is however not exanded on in the EDR literature.

For 'centre of relation' concepts, supplementary information can be included in like vein, however the available EDR description of this information is vague. What is clear however is that e.g. events of movement can have 'property information' associated wtih them, e.g. 'spatial relation' and also 'phase' information, e.g. for <approach> the phase would be given as 'shorten distance'.

In summary, we can say that the EDR Concept Dictionary provides an ontology of interlingual concepts. This ontology is organised by conceptual case relations and semantic relations (the latter yielding further an IS-A network). There is apparently nothing particularly innovative about the ontology – it implements many features to be found in classical AI knowledge bases. What is of interest is the EDR methodology for arriving at a set of interlingual concepts.

The following general points can be made, given the available documentation: There is a vague boundary between more simple and more complex concepts: EDR gives the impression that it is interested in describing highly complex concepts that approach the meanings of full sentences. Thus, it would seem that EDR is interested in building an entire knowledge base, which goes much further than relating words to their concepts with a measure of classification and inter-relation. However, EDR did not expect to complete full embedded CRR descriptions, but that it would complete as far as possible the recording of basic concept entries (i.e. relations between headconcepts). It appears to be difficult to conceive of a methodology that would allow reasonable unification of highly complex concepts representing full sentences to yield a set of *interlingual* concepts. So far, EDR has, as far as we know, evolved a methodology for unifying only concepts related to individual words (or short phrase, idioms, compound words), to yield an interlingual set of concepts. There is no mention of default values for various properties. Defaults have been generally found useful in other projects we know of. There appears to be no information recorded on specific values (e.g. that a car has 4 wheels, or not more than 4 wheels). There appears to be no attempt at incorporating relaxation of preference (i.e. that <drink> typically prefers an animate agent but not always). It is possible that exhaustive corpus-based work will yield instances of e.g. < drink> being used with an inanimate agent. It would appear that this fact could not be easily *related*

to a *typical* use of <drink>, in the EDR design. Equally, there does not appear to be any information recorded on the relative significance of various elements of a complex concept (e.g. that a car must significantly have wheels but need not have a radio). There appears to be no consideration of the role of scalar attributes to reduce the complexity of the ontology. e.g. <child> and <adult> need not be represented in the ontology as separate concepts, but could be incorporated in the concept < human>, and an age range recorded to distinguish varieties of human by age.

It is not at all clear that the Concept Dictionary will be reusable in in its entirety in a meaningful sense. In the absence of detailed information no firm judgement can be formed. However, we note the following: The list of basic headconcepts will probably provide as reasonable a set of basic interlingual concepts as any other project. It is too early to say whether this set will however be in any sense 'better' than those of many other projects which have elaborated only small ontologies. There is some room to doubt the advisability of elaborating interlingual concepts according to the methodology espoused by EDR. The bulk of the complex CRRs would, in all probability, be useful only in certain situations. There is a great deal of doubt in the field in general as to how to represent complex conceptual meaning. There appears to be a number of elements missing (but found in other well-known knowledge bases) that would render the complex CRRs more useful (see above). As the complex CRRs are effectively built up by recording meanings of a relatively small number of corpus sentences (20 million per language), it appears, at least on the surface, that there is little room for allowing for flexibility of interpretation when the result is applied in the interpretation of new sentences and expressions, at least for general language.

3.3.3.7 Synoptic table of the information types in the EDR dictionaries

Table 15: Information types in the EDR dictionaries

	Entry component		Present	Information content
1	Headword		✓	Text form, non-linguistic stem, compound constituents, string with syllable markers / uninflected part in katakana (used for kana-kanji conversion)
2	Phonetic transcription		✓	IPA / Katakana
3	Variant form		✓	Separate entry
4	Inflected form		✓	
5	Cross-reference		✓	Via headconcept relation
6	**Morphosyntactic Information**			
	a	Part-of-speech marker	✓	
	b	Inflectional class	✓	In extenso, adjacency information also (inflectional information distributed over several levels)
	c	Derivation	✓	Minimal. Also adjacency information (derivational information distributed over several levels)
	d	Gender		Information about the gender of the entry in SL and TL
	e	Number	✓	Information about the grammatical number of the entry in SL and TL
	f	Mass vs. Count	✓	Special treatment of nouns in number agreement
	g	Gradation	✓	
7	Subdivision counter			Not explicitly. Concept reference used (concept classification)
8	Entry subdivision			Not explicitly. Concept reference used (concept classification)
9	Sense indicator			Not explicitly. Concept reference used (concept relation, concept classification)
10	linguistic label			✓
11	**Syntactic Information**			

	a	Subcategorization frame	✓	(i.) Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.
	b	Obligatority of complements	✓	
	c	Auxiliary	✓	
	d	Light or support verb construction	✓	
	e	Periphrastic constructions	✓	
	f	Phrasal verbs	✓	
	g	Collocator	✓	(i.) typical subject /object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented (iii.) cooccurrence information
	h	Alternations	✓	
12	**Semantic Information**			
	a	Semantic type	✓	
	b	Argument structure	✓	
	c	Semantic relations	✓	
	d	Regular polysemy	✓	
	e	Domain	✓	Separate terminological dictionaries
	f	Decomposition	✓	
13	Translation		✓	
14	Gloss		✓	
15	Near-equivalent		✓	
16	Example phrase (straightforward)		✓	Link to corpus
17	Example phrase (problematic)		✓	Link to corpus
18	Multiword unit		✓	
19	Subheadword *also* secondary headword		✓	Via concept reference
20	usage note		✓	
21	Frequency		✓	Based on corpus

3.3.4 SYSTRAN

SYSTRAN uses two dictionaries:

1. "Stem Dictionary" containing single words with grammatical information and translations

2. "Expression Dictionary" for all multiple word expressions and for rule-based expressions. These range from simple noun compounds to complex lexically driven rules.

Based on syntactic and semantic information in the Stem Dictionary, the SYSTRAN parser attaches information on the syntactic function of the word in a given sentence and sets syntactic relationships between words. This information can be checked in the rules written in the Expression Dictionary.

For more information on these dictionaries see (Gerber and Yang, 1997).

The SYSTRAN example entries given in chapter 5 illustrate rules from the "Expression Dictionary", unless otherwise indicated.

3.3.5 Lexical Conceptual Structure Lexicons

The aim of the translation system developed at UMIACS is to generate natural language sentences from an interlingual representation, the Lexical Conceptual Structure (LCS). This system has been developed as part of a Chinese-English Machine Translation system, however, it promises to be useful for many other MT language pairs.

The generation system has also been used in Cross-language information retrieval research (Levow et al., 2000).

Lexical Conceptual Structure is a compositional abstraction with language-independent properties that transcend structural idiosyncrasies. This representation have been used as the interlingua of several projects such as UNITRAN (Dorr et al., 1995) and MILT (Dorr, 1997).

An LCS is a directed graph with a root. Each node is associated with certain information, including a *type*, a *primitive* and a *field.*

The type of an LCS node is one of *Event*, *State*, *Path*, *Manner*, *Property* or *Thing*.

There are two general classes of primitives: closed class or structural primitives (e. G., CAUSE, GO, BE, TO) and open class primitives or constants (e. g., REDUCE+ED, TEXTILE+, SLASH+INGLY).

Suffixes such AS +, +ED, +INGLY are markers of the open class of primitives. Examples of fields include LOCATION, POSSESSIONAL, IDENTIFICATIONAL.

An LCS captures the semantic of a lexical item through a combination of semantic structure (specified by the shape of the graph and its structural primitives and fields) and semantic content (specified through constants).

In this way, for example, the semantic structure of a verb is something the verb inherits from its Levin verb class whereas the content comes for the specific verb itself. So, all the verbs in the “Cut Verbs-Change of State” class have the same semantic structure but vary in their semantic content (for example, chip, cut, saw, scrape, slash and scratch).

3.3.6 Microsoft Bilingual Resources

Microsoft current English lexicon consists of data acquired from two MRDs: LDOCE and AHD (3rd ed.). The main lexicon can be thought of as a repository of all the sense distinctions for the headwords in those two dictionaries. Each sense is assigned to a distinct record under each headword; additionally, undefined run-ons and "irregular" inflected forms are promoted to full entry status, of course maintaining bidirectional links between the new records and the parent ones. So, "wept" links to "weep", while "weep" lists "wept" as one of its inflected forms.

The overall architecture of the lexicon is to extract as much information as possible out of existing resources, ranging from raw MRD data to dictionary definitions to full text corpora. The information extracted is then folded right back into the dictionary, for use by the various dictionary clients (morphology, syntax, logical form rules, translation, generation, etc.), and to bootstrap further dictionary work. For example, one of the first things done was apply the derivational morphology rules automatically to each headword in the dictionary, which allows the identification of the bases of lexicalized derived forms lacking explicit links in the MRDs. That in turn allows the linking of all forms in the same derivational paradigm. As a result, the dictionary stores the information that the words 'belief, believe, believer, disbelieve, believable, believably, unbelievable, unbeliebably' are all part of the same derivational paradigm; that in turn can be useful during generation.

There are at least four sets of secondary, derived lexicons created in this fashion, and which are stored in the dictionary file system: (a) morphological lexicon, which has been used as a stand-alone lexicon for some applications; (b) syntactic lexicon, which orders entries by part of speech, packing ambiguity internal to a part of speech inside each entry, since the grammar is very flexible and does not attempt disambiguation beyond part of speech; (c) monolingual MindNet, created by parsing definitions, resulting in a rich network of relations between words, which then can be used to compute similarity between headwords; (d) bilingual MindNet, which stores parallel bilingual fragments learned by processing aligned bilingual corpora. Bilingual lexica are also used, which however store simple word correspondences, and are used primarily while constructing the bilingual MindNet, and as a repository of default translations should a translation not be found in the bilingual MindNet.

Because these derived lexica are all created dynamically, by applying morphological, syntactic, or logical form rules to the input definitions or corpora, they can be rebuilt automatically in a very short time (ranging from a few minutes to a few hours); consequently, the data in those dictionaries continues to improve as the rule bases are improved.

Not all the lexical maintenance work is automatic. In addition to thought and experimentation in trying to come up with techniques that can mine data for more lexical information, there is quite a bit of manual maintenance involved in making sure the dictionaries contain the right information for all their clients. However, the information typically added is morphological or syntactic and very rarely new senses are manually added to the dictionary. Because the syntactic grammars are ever-evolving, the data that goes into the dictionary will also evolve. It is one of the jobs of the lexicographer to know the "internal landscape" of the dictionary, so that inconsistencies and unnecessary redundancies can be avoided; but the data of the syntactic lexicon, for example, will be as rich as each grammar needs it to be.

Table 16. Information types in Microsoft bilingual resources

	Entry component		Present	Information content
1	Headword		✓	Lemma, canonical form of capitalization (so "Polish" and "polish" are different headwords)
2	Phonetic transcription		✓	Uses ARPABET for English, as well as AHD's native scheme.
3	variant form		✓	cross referenced
4	inflected form		✓	Irregular forms lexicalized, regular forms handled by morphological rules—effectively all forms may be accessed in lexicon, both for analysis and generation.
5	Cross-reference		✓	Same as in LDOCE and AHD
6	**Morphosyntactic Information**			
	a	Part-of-speech marker	✓	11 possible: Noun, Verb, Adj, Adj, Conj, Prep, Pron, Ij, Posp (postposition), Funcw (function word for particles in Asian languages), Char (for punctuation characters) POS further subcategorized with additional features (so determiners are Adj with subcat Det)
	b	Inflectional class	✓	paradigm marked for each word for each part of speech
	c	Derivation	✓	complete and cross-linked
	d	Gender	✓	grammatical gender marked
	e	Number	✓	
	f	Mass vs. Count	✓	
	g	Gradation	✓	
	#H	Collectives	✓	
7	Subdivision counter		✓	Sense distinctions maintained from source MRDs (LDOCE and AHD), but no inherent sense hierarchy in our system
8	Entry subdivision		✓	lexemes may be differentiated within a part of speech record in the syntactic lexicon. (MS dictionary is very dynamic in nature; may have one static form, but be accessed in logically different ways)
9	Sense indicator		✓	Only those found in definitions from MRDs
10	linguistic label		✓	domain, style, etc. indicators from MRDs
11	**Syntactic Information**			
	a	Subcategorization frame	✓	superset of LDOCE codes
	b	Obligatority of complements		No (contains LDOCE codes, but use is not strictly enforced)
	c	Auxiliary	✓	
	d	Light or support verb construction		
	e	Periphrastic constructions		No, other than what is in MRD definitions

	f	Phrasal verbs	✓	with syntactic subcategorization and distinctions between prepositions and adverbials
	g	Collocator	✓	extracted from MRD definitions and then used during parsing
	h	Alternations	✓	many but not complete
12	**Semantic Information**			
	a	Semantic type		(no ontology planned)
	b	Argument structure		No (for now—we're working on it)
	c	Semantic relations	✓	Currently 26+ relations in MindNet, some of which follow: Attribute Goal Possessor Cause Hypernym Purpose Co-Agent Location Size Color Manner Source Deep_Object Material Subclass Deep_Subject Means Synonym Domain Modifier Time Equivalent Part User
	d	Regular polysemy		No
	e	Domain	✓	from MRDs
	f	Decomposition		No
13	Translation		✓	both from Bilingual MRDs and learned from aligned corpora
14	Gloss			No
15	Near-equivalent		✓	Generally no, but may be learned from aligned corpora
16	example phrase (straightforward)		✓	many.
17	Example phrase (problematic)		✓	Treated the same as in (16); many, many examples from aligned corpora
18	multiword unit		✓	Fixed multiword units as well as phrasal verbs can be lexicalized (English lexicon contains about 28,000 such). Others, such as idioms, are learned from aligned corpora (see examples from Grishman/Palmer below).
19	subheadword *also* secondary headword		✓	Yes – become linked headwords
20	usage note		✓	Those from MRDs
21	Frequency		✓	POS frequencies and semantic relation frequencies

3.3.7 Lexicography for speech-to-speech translation: VerbMobil

3.3.7.1 Application requirements

The complex of problems facing the lexicographer in speech-to-speech is well illustrated by lexicography in the Verbmobil project, which terminated in September 2000 after 8 years of funding (not counting a smaller two-year pilot phase known as "ASL - Architectures for Speech and Language"). The global goal of the Verbmobil project was to develop a prototype for portable speech-to-speech translation systems; as awareness of the magnitude of the problem grew, so did the power of the hardware and the sophistication of software modules and their interaction, so that the goal was attained on the basis of a high-end laptop computer as well as in a server and mobile phone environment.

(Wahlster, 2000) contains the most comprehensive published documentation of the Verbmobil project. The contributions by Gibbon & Lüngen (lexicography), Emele & al. (transfer), and Burger & al. (spoken language corpus annotation), are particularly relevant to the lexicographic work in Verbmobil, but several other chapters are also relevant in various ways. Many results of the earlier phases of lexicography in the Verbmobil project are represented in previous documents of the EU funded EAGLES project, including (Gibbon & al., 1997) and (Gibbon & al., 2000). References to Verbmobil technical reports are not given, as these are too numerous to be justified in an overview of this kind, and can easily be consulted via the literature mentioned here.

This overview concentrates mainly on the new problem of spoken language lexicography with which the Verbmobil project was confronted, rather than on machine translation lexicography. There are also many non-lexicographic aspects of spoken language translation which cannot be covered here, such as the highly elliptical and ambiguous character of spoken language, recovery from fragmentation, re-starts, errors, hesitations, the translation of prosody and speaker attitude in culturally different environments, the adaptation of voice output to speaker input.

3.3.7.2 Problems of spoken language lexicography

A wide range of logistical and module-specific subordinate goals were pursued in the Verbmobil project, the most conspicuous of which was the novel problem of handling elicited but largely spontaneous spoken dialogue at all levels. In terms of lexicographic domains, this resulted in the birth of a new sub-discipline of spoken language lexicography, in which traditional lexicographic information types (morphological, syntactic, compositional semantic, domain semantic, pragmatic) were combined with machine translation information (bilingual transfer information) and with additional information about the pronunciation and modulation of spoken language: phoneme patterns, enhanced with prosodic information such as syllable boundary and stress marking, pronunciation variants, lexicalised discourse phenomena such as hesitation markers.

The new, heterogeneous set of lexicographic subdomains engendered a new lexicographic methodology: each subdomain was not only developed by linguistic experts from disciplines with very different substantive, terminological and methodological backgrounds, but at every level linguistics rapidly turned into formal and computational linguistics as cooperation with speech engineering research units and specialists from theoretical computer science and software engineering developed.

Consequently, the major lexicographic problem for most of the project lifetime was heterogeneity:

1. of input to the lexicon acquisition and integration process,
2. of access and output to the databases required for each component, and
3. of versions due to the decentralised development of lexical information.

As the size and complexity of the lexicon grew, it became clear that new coordination techniques were required. A number of measures were introduced in order to cope, including a clear distinction between offline lexicon and online lexicon (system component lexicon), standardisation and automatic validation of the spoken language transcriptions on which the lexicon was based.

Offline lexicon: The offline lexicon was to integrate as many types of lexical information as possible, in as procedurally neutral a fashion as possible, in the form of a coherent but easily extendible and accessible database. The simplest possible solution was adopted: a classical UNIX database with ASCII encoded field contents and separators. This simple database type was easily processable by engineers, computational linguists and computer scientists. Conventions for structured fields (disjunctions), field and record separators were introduced, and statistics on the current state of field-filling were maintained. The database was distributed initially on the internet by ftp. In 1994 a WWW client concept for lexical access ("HyprLex") was introduced and further developed until the end of the project. Associated with the offline lexicon were additional innovative functionalities such as search of lexical class on the basis of parametrisable phonetic similarities, and a transcription concordance. The use of the WWW introduced a new quality of interaction, and supported database consistency in a novel way by permitting all project users to access a single token of the database.

Online lexicon: The system component lexica were de-centrally developed, and thus responsibility of the system component builders. The system component builders supplied the lexicographic integration team with examples of the lexicon formats they required, and in many cases also their own lexica containing specialised phonological, syntactic, semantic and transfer information. This input was re-formatted (in some cases reverse engineering was necessary) for integration, and output from the offline lexicon was provided, either in the standardised format already noted, or in the formats required by the different component developer groups.

Spoken language corpus lexicography: Lexicographic work in the Verbmobil project was necessarily (almost) exclusively corpus lexicography based on orthographic transcriptions of digitally recorded spoken language appointment scheduling dialogues. Two main varieties of exception to the corpus-based methodology were needed: first, completion of morphological paradigms; second, completion of semantic paradigms with accidental gaps, e.g. names of days, weeks, months. The corpus orientation, as opposed to introspective vocabulary selection, was mandatory for methodological reasons: the statistical training methods required for speech recognition demand actual corpus data.

Transcription validation: A condition on the transcription corpus which was the basis of the lexicographic work was absolute consistency. Previous experience with transcription had revealed many possible sources of transcriber inconsistency, and the error proportion of conventionally produced lexica such as CELEX is much too high for speech technology applications.

Criteria were introduced by the corpus creation groups in order to ensure consistency:

1. Use of canonical phonemic transcription, not phonetically detailed or impressionistic conversational transcription.

2. Development of encodings for a number of classes of spontaneous speech phenomena (fragmented words, hesitations)

3. Development of encodings for a number of classes of non-speech sounds.

4. Provision for comments.

In addition, the lexicography team designed and implemented a parametrised transcription checker trlfilter with two functions:

1. Error checking (spell checking) of the transcription input in order to ensure consistency;
2. Re-formatting of the input into alternative formats required by speech recognition teams.

Automatic paradigm completion: Part of the lexicographic team was a morphology unit which developed a full description of inflexional morphology for spoken German with an inheritance hierarchy of generalisations over inflexional class and subclasses based not on orthography (as in all previous similar lexica) but on phonological and prosodic generalisations. The paradigm generator based on this morphological model ensured consistent generation of all inflected forms and their correct morphological categories.

Quality control: The ultimate, though indirect, lexical quality control criterion in the lexicon evaluation process was the quantitative performance of each component of the Verbmobil system and the translation performance of the system as a whole. In particular the statistical training methods used for speech recognition meant that inconsistencies would have immediate and possibly quite disastrous results on speech recognition rates.

3.3.7.3 Lexical coverage

A major difference between lexicography for spoken language (in the sense of spoken language systems) and lexicography for written language lies in the absolute size of lexical coverage. At an early stage of the Verbmobil project, the lexicography team introduced a distinction between extensional coverage, i.e. the number of lexical objects (entries) included in the lexicon, and intensional coverage, the number of properties associated with lexical objects. In database terms, extensional coverage amounts to the number of records, intensional coverage amounts to the number of fields.

3.3.7.4 Multilingual extensional coverage

By the standards of written language text corpora, the absolute coverage is rather small, for three very good reasons. First, The Verbmobil lexicon had to be almost totally corpus-dependent. Second, spoken language corpora are highly complex signal databases which are extremely labour-intensive to process; reliable cross-checked transcriptions may take several hundred times real time to produce, i.e. an hour of recording may take several hundred hours of transcription production and checking time. Third, the corpora are always very specifically task-oriented and are constructed as required, because speech recogniser training does not easily generalise from one corpus to another.

The criterial definition of vocabulary coverage in the Verbmobil lexicon is very simple:

The extensional coverage of the lexicon is the set of labels of edges in the word hypothesis graph at the interface between the speech recognisers and the parsers.

This vocabulary is derived from the corpus, and contains inflected forms, non-inflecting words, and representations of discourse particles such as hesitation phenomena, and of noises. Anything which is not in this set is an "out of vocabulary item" (OOV item).

The notion of translationally equivalent wordlist was introduced in order to define the wordlists for English and Japanese:

The translation equivalent of a given wordlist WL extracted from a dialog corpus C is the list of words of the target language that are needed for the translation of C.

This definition was operationalised with reference to translation transfer rules:

> *The translation equivalent of a wordlist WL, extracted from a dialog corpus C, is the list of lemmata that occur on the right hand side of a transfer rule T, whose left hand side contains a semantic lemma with a morphologically corresponding entry in WL.*

Towards the end of the Verbmobil project, the corpus included about 25000 dialogue turns, and 10000 words. The notion of "word" in this kind of corpus requires immediate clarification. For speech recognition, the measure is in terms of fully inflected forms of words, as a speech recogniser literally requires a surface form to match to the signal not an underlying lemma. But in a corpus of this size, the number of lemmata which can be extracted is not very different from the number of fully inflected forms; by the standards of written language text corpora, in a corpus of this size a large number of hapax legomena would be expected. The full set of inflected forms projected from this basic corpus set came to over 50000; a ratio of approximately 1:5 for stems to inflected forms has frequently been observed for German.

A number of extensions to the basic corpus lexicon were made; for the speech recognition systems this meant developing techniques of OOV word recognition.

3.3.7.5 Intensional coverage for German

The main constraint on Verbmobil intensional coverage was quite unlike that found in many types of written language lexicography: the types of lexical information were dictated by the system architecture, which was to some extent evidently determined by linguistic considerations, but mainly by considerations of feasibility and experimentation with new techniques. The architecture permitted alternative speech recognisers to be plugged in, a prosodic component, a morphological component (in the first version), alternative parsers, a compositional semantic component, a domain modelling component, a transfer component and alternative speech synthesisers. Each of these made different and in many cases rather unrelated demands. However, the criterion for integration conformed exclusively to the definition of extensional coverage: the entries were all associated with the forms attested in the corpus. This meant that a number of satellite lexica, in which other forms of lexical organisation needed to be derived from the main lexicon, particularly for syntactic parsing, semantics, and transfer, based on classical lemma or concept definitions. However, these were then re-integrated into the main lexicon by the lexicography team.

Fig. 15: Early web interface to Verbmobil lexical database.

The most straightforward way to illustrate the intensional coverage is by example. Figure 15 shows the 1996 web interface to the lexical database, with filter buttons for microstructure elements. The following output is from a query to this interface, with output for each type of lexical information. The 1996 interface is selected because in the second Verbmobil phase the extent of lexicographic coordination work was drastically reduced in view of the existing available work and techniques, and revised prototype oriented goals. All versions of the interface can be consulted directly at: http://coral.lili.uni-bielefeld.de/VM-HyprLex/ .

```
VM-HyprLex results

Server: coral.lili.uni-bielefeld.de (via tmp.430.html)
Date: Tue Feb 27 22:46:48 CET 2001
Specification: String / Key / All / bielefeld.lexdb.v3.3

Number of matches = 1

Entry 2537 matches String key Terminabsprache:
  BIorth:       Terminabsprache
  BIorthseg:    Termin#ab#sprach#+e
  BImorpro:     tE6.m'i:n#?''ap#Spr''a:.x#+@
  BIorthstem:   Termin#ab#sprach
  BIphonstem:   tE6.m'i:n#?''ap#Spr''a:x
  BIflex:       N,akk,sg,fem
```

```
                N,dat,sg,fem
                N,gen,sg,fem
                N,nom,sg,fem
BIlemma:        Terminabsprache
BIspell:        --
BIproper:       --
BIcompsem:      ObjEreig
BICD1:          cd1=2_cd12=7_cd3=2_cd4=3_cd5=1
BICDall:        15
BIpercent:      0.00568005%
BIrank:         977
BIortherror:    Termin-Absprache,-
BLAUBEU:        --
DemoWL:         demo-wl
RQH-WL:         --
BIhitlist:      hit#977=15
FPWL3:          fpwl
KIcanon:        tE6m'i:n#Q"ap#Spr"a:x@
KIfreq:         14
IMSlem:         Terminabsprache
IMSpos:         NN
IMSfreq:        8
SIEMENSorth:    Terminabsprache
SIEMENScats:    sem_lex(nr,terminabsprache)&
                nr:rel=terminabsprache&
                sortal_Terminabsprache(nr)&
                count_noun_norm(nr)&
                subst_klasse2_1(nr)
                terminabsprache&
                sortal_einigen_auf&
                count_noun_norm&
                subst_klasse2_1
SIHUBval:       --
BIgloss:        appointment_scheduling
IBMorth:        --
IBMmorph:       --
IBMHUBsyn:      [gender:fem,
                 number:sg,case:ncase_v,
                 syn_ibm:[phon:'Terminabsprache',
                          cuf_macro:common_noun_syn],
                 person:3]
TUBsem:         terminabsprache_&_communicating_&_-
TUEBcomp:       terminabsprache:
                  compound(terminwoche,
                           first(termin),
                           second(absprache),
                           semrel(arg3_rel)).
IMSrule:        terminabsprache:
                [H: terminabsprache(I)]
                <->
                [H:scheduling(I),
                 H1:indef(Y,H2),
                 H2:appointment(Y),
                 H3:of(I,Y)].
```

The definitions of the microstructure elements are as follows (sources in parentheses).

- Orthography, according to Verbmobil orthographic conventions (Daniela Steinbrecher & Dafydd Gibbon, Bielefeld).
- Segmented orthography (Doris Bleiching & Daniela Steinbrecher).
- Morphoprosodic transcription, with accentual word prosodic marking (single quote for primary stress, two single quotes for secondary stress), and segmentation on two levels:

morph segmentation and syllable segmentation. The phonemic symbols in the transcription correspond to standard international SAMPA conventions (Doris Bleiching & Daniela Steinbrecher, Bielefeld).

- Orthographic stem (Doris Bleiching, Bielefeld).
- Morphoprosodic stem (Doris Bleiching, Bielefeld).
- Inflexion categories are represented as a vector containing ordered information about the (morphological) part of speech and values of inflexional attributes such as case and number (Doris Bleiching & Guido Drexel, Bielefeld).
- Frequency and rank information, 4 fields (Dafydd Gibbon, Bielefeld).
- Orthographic errors: Orthographic errors automatically percolate into the LexDB because processing is automatic; they are checked with a standard orthography list after integration. The orthographic error list is made available for list checking and for correction by VERBMOBIL partners (Dafydd Gibbon, Bielefeld).
- Corpus source information, 5 fields (Dafydd Gibbon, Bielefeld).
- IMS POS tags: The tags assigned to tokens in the CD-ROM corpus by the IMS Stuttgart stochastic tagger (Martin Emele, Stuttgart).
- IMS POS frequencies: The frequencies of occurrence of an item as a specific part of speech as assigned by the IMS Stuttgart stochastic tagger, and the sum of these frequencies (Martin Emele, Stuttgart).
- TP 14 canonical phonemic transcription: The canonical corpus transcription used in TP 14 (IPK Kiel).
- TP 14 frequencies:Frequencies for token occurrences of items in the Kiel canonical phonemic word list in the transliterations processed by the IPK Kiel filter (IPK Kiel).
- Information from the Siemens parser group, 2 fields (Hans-Ulrich Block, Siemens).
- Information from the IBM parser group, 2 fields (Anke Feldhaus, IBM).
- Spelling compounds, of two main kinds: first, the standard abbreviation or acronym, and second, the uptake spelling, or spell-out, in which a word is spelt letter by letter for the sake of clarity (Dafydd Gibbon, Bielefeld).
- Proper names: These are annotated separately as they play a role in the selection of the Research Prototype Word List (Dafydd Gibbon, Bielefeld).
- Morphosemantics for compounds: The macros for the morphosemantics of compound words define constraints for the morphological component of the VERBMOBIL Research Prototype (Harald Lüngen & Kerstin Fischer, Bielefeld).
- Verb valencies: Valency structures for verbs, including some function verb syntagmas (`Funktionsverbgefüge'), based on the `arg1, ... , argn' model (Johannes Heinecke, Berlin).
- English glossary: English glossary for text-to-speech single word translation in the VERBMOBIL Research Prototype (Dafydd Gibbon, Bielefeld).
- CUF syntactic categories: Lexical syntactic categories in the CUF unification formalism (Johannes Heinecke, Berlin).
- Stuttgart transfer database: IMS Stuttgart database for transfer component, containing corpus tags, glosses, transfer rule information (Martin Emele, Stuttgart).
- Semantic evaluation: TU Berlin semantic evaluation relations (Joachim Quantz, Berlin).
- Tübingen compound noun semantics: Transfer relevant semantics for nominal compounds with TUEB orthographic keys (Sabine Reinhard, Tübingen).
- Stuttgart transfer rules: Lexical transfer rules, with IMS orthographic keys (Martin Emele, Stuttgart).

The combined external and internal coverage statistics were used for evaluating lexicographic progress, as shown below:

```
Coverage figures for bielefeld.lexdb.v3.3
Generated by gibbon with ./dbstats
Mon Mar 18 22:29:47 MET 1996
1. BIorth          8081 100.00%
2. BIorthseg       7577  93.76%
3. BImorpro        7577  93.76%
4. BIorthstem      7577  93.76%
5. BIphonstem      7577  93.76%
6. BIflex          7577  93.76%
7. BIlemma         7577  93.76%
8. BIspell          246   3.04%
9. BIproper         517   6.40%
10. BIcompsem        139   1.72%
11. BICD1           5851  72.40%
12. BICDall         5851  72.40%
13. BIpercent       5851  72.40%
14. BIrank          5851  72.40%
15. BIortherror      406   5.02%
16. BLAUBEU          508   6.29%
17. DemoWL          1292  15.99%
18. RQH-WL           562   6.95%
19. BIhitlist       1000  12.37%
20. FPWL3           2461  30.45%
21. KIcanon         5404  66.87%
22. KIfreq          5404  66.87%
23. IMSlem          2288  28.31%
24. IMSpos          2288  28.31%
25. IMSfreq         2288  28.31%
26. SIEMENSorth     3267  40.43%
27. SIEMENScats     3267  40.43%
28. SIHUBval          71   0.88%
29. BIgloss         2280  28.21%
30. IBMorth          390   4.83%
31. IBMmorph         390   4.83%
32. IBMHUBsyn       1773  21.94%
33. TUBsem           174   2.15%
34. TUEBcomp          19   0.24%
35. IMSrule          852  10.54%
Number of records:     8081
Fields per record:       35
Number of fields:    282835
Fields filled:       114233
Overall coverage:        40%
```

3.3.7.6 Lessons for spoken language lexicography logistics

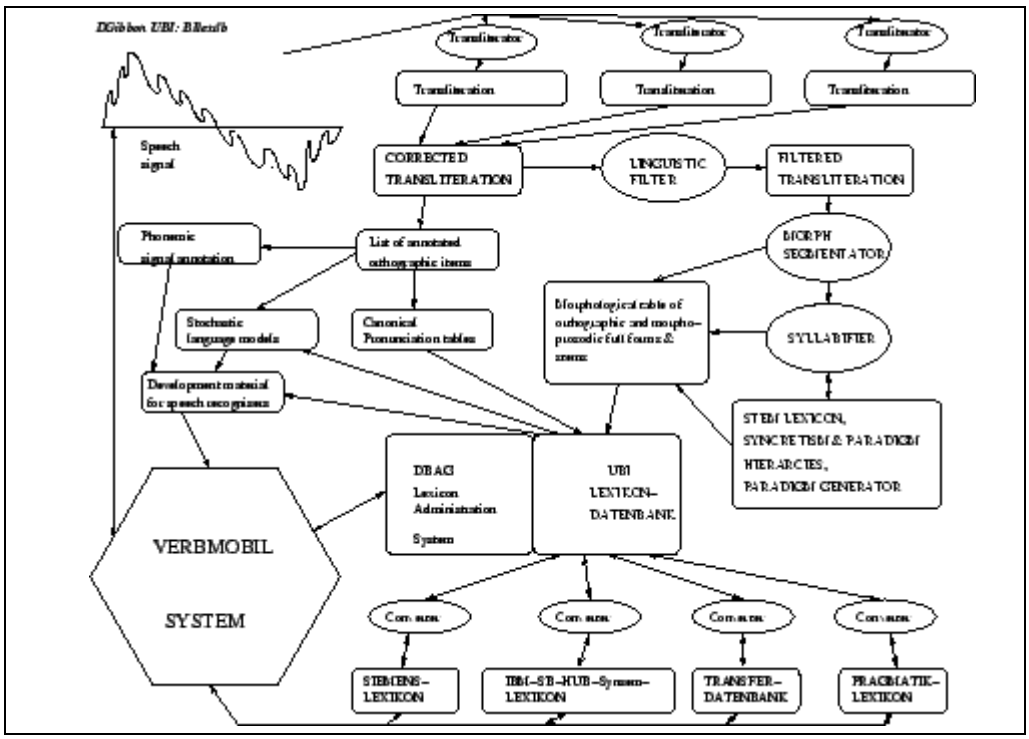


Fig. 16: Lexicographic workflow in Verbmobil.

The lexicographic workflow in the Verbmobil project, on which the collation and integration of lexical information was based, is shown in fig. 19. The transcription process is shown at the top, proceeding through the transcription verification and lemmatisation processes to grapheme-phoneme conversion, prosodic enhancement, morphological paradigm completion, to the provision of the lexical database for system developers.

It is clear that in a new, hybrid and experimental software and lingware development environment on the scale of Verbmobil, in which many components were designed to be competitive alternatives, the lexicographic development strategy had to be adapted as needs grew and technological possibilities opened up. A uniform theoretical basis for lexical information, and indeed uniform formatting conventions were not possible. Consequently, a pragmatically designed database prototype was developed in the early stages, in close consultation between all lexicographic contributors and users, and provided stable service throughout the project. More detailed on the relation of lexicography to other aspects of system development in the Verbmobil project (Wahlster, 2000) should be consulted.

One result of lexicographic development in the Verbmobil project was to lay out clearly the requirements for future work in spoken language system lexicography. Subsequent projects world-wide have benefited not only from the lexical content, but also from the software and the overall coordination methodology developed for the distributed development environment of Verbmobil.

3.3.8 GENELEX

GENELEX defines a generic model for lexicons, theory and application independent, as being based on EAGLES work on the lexicon, and due to the fact that EAGLES recommendations and GENELEX model have been established after consulting and generalizing a number of theories and existing NLP lexicons, as well as after identifying different users' needs depending on the kind of applications: text tagging or analysing, generation, automatic indexing, assisted translation, NL query to database. It is designed to ensure that application dependant models of data and applicative dictionaries can be derived from this repository of information, by mapping the application model from the generic one.

It grounds the specifications on Entity/Relationship for conceptual modeling and an SGML DTD (instantiated for each language) as formal specification and as a reference format of interchange, in particular for generic tools (extended GENELEX tools). Additional constraints (for each language) have to be specified and verified by dedicated tools (extended GENELEX tools) associated to the lexicographic work-stations that will be developed and reused on the base of the common core.

GENELEX is designed to fulfill the needs of a wide range of NLP applications representing different kind of information in an integrated and coherent model without committing towards a given linguistic theory. A lexicon conformant to this model is not an application lexicon, but contains the basic information needed by applications. Applications can extract the required data in the application format.

This presupposes a high level of precision in the description, so that these bases can be independent from the applications. It also presumes that the available information is self-sufficient and fully explicit, and, at the difference with dictionaries for human readers, does not require human interpretation or non-explicit knowledge. The model allows variable granularity of information, and the encoding of basic information can be performed within this model at one step, and its refinement in another step as incremental information.

The model is designed as a whole: it accounts for basic levels of linguistic description (orthography, inflections, morphosyntax and minimal syntax as subcategorization) and also for more refined information as derivation at the morphological level, refined syntax, lexical semantics and mutilingual links based on syntax and semantics levels.

The requirement of expliciteness and variability of granularity is fulfilled by a descriptive structure where different descriptive elements interact and some complex ones are described by more basic ones, themselves described by smaller ones; all these descriptive elements of different levels are identified as such, and can be shared by the descriptive elements of various others of higher level: linguistic generalities are captured at different levels. The model can be seen as containing both the "traditional" linguistic level of elements of description to be attached to lexical entities and the explicitation by analytic description of those "traditional" elements referred. For instance, it is not enough to give the class of the inflectional paradigm associated to an entry, and it is important and necessary that the model gives means to explicitly describe it.

Three descriptive levels: morphology, syntax and semantics, have been described independently and coherently connected the one to the other: this guarantees the possibility to encode a certain descriptive level without taking into account the criteria of another level. It permits to distinguish different syntactic behaviours on pure syntax criteria, and independently of the fact they share the same meaning (Semantic Unit) or not. It permits to refine the description of one level (i.e. syntax or semantics) without changing the description of others.

This architecture is one answer to the requisite of genericity and expliciteness, it doesn't mean in absolute that applications need to have the same approach to the structure of the lexical data: they

may identify two levels (for instance morphological and syntactico-semantic), or just one flat level carrying information from the three original levels. Depending on the choices of an application at a given time, dedicated mappers will have to be written to extract the right information in the right structure and the right format.

3.3.8.1 The GENELEX architecture

The global architecture of the lexicon is as follows:

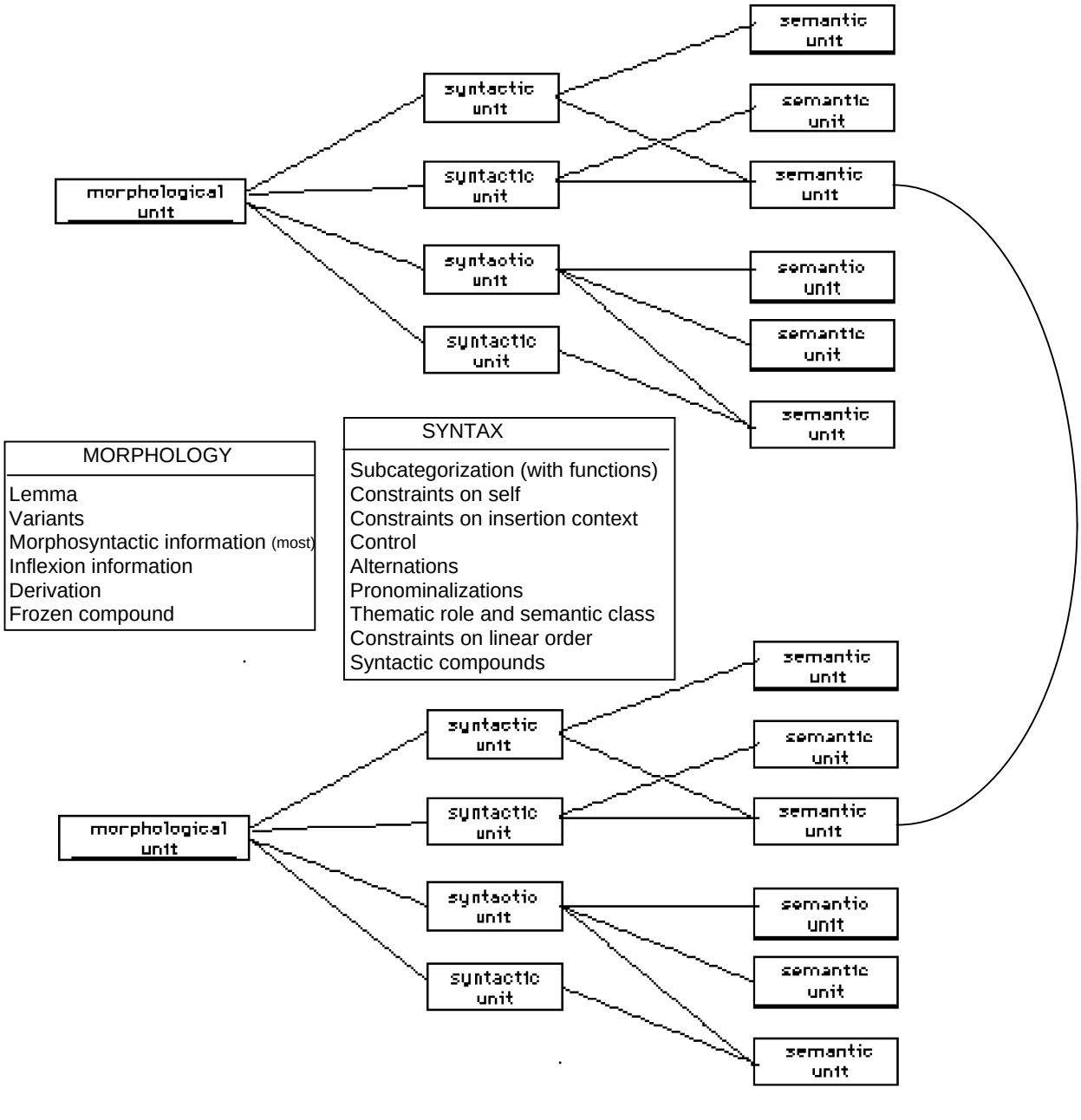


Fig. 17: global architecture of theGENELEX lexicon

There is no entity as a "lexical unit", but depending on the point of view, they can be reconstructed: a complete monolingual lexical entry can be seen either as a progression through the three levels,

either as the whole set of information accessed by the morphological unit (as for editorial dictionaries), either as the set of information accessed through one semantic unit, or as a set of information regarding one level. It is important to notice that a syntactic unit has access to the morphological information it is associated to through the morphological unit, as well as a semantic unit is associated to one (or more) syntactic contexts through the syntactic unit. As a consequence of this structure, the distinctions in each level are made with criteria of this level: for instance, there is no formal need for distinctions related to polysemy until the semantic level is described, and thus morphological and syntactic information can be shared.

It has to be noted that multilingual links operate mainly at semantic level, but given the fact that semantic units are always associated to syntactic units, implicitely, syntactic units are linked when associated to linked semantic units.

3.3.8.1.1 Morphology

The morphological level is where all the information about the form of the lexical entry can be found: what is related to orthography, inflections and variations, derivations, affixes, internal composition of frozen compounds. The morphological unit (MU), which corresponds intuitively to the lemma of traditional dictionaries entries, represents an equivalence class of related forms associated to different information(see above). The set of possible labels for morphosyntactic categories and subcategories, (catgram and subcatgram attributes) as well as relevant inflectional features and values is based on EAGLES recommendations for morphosyntactic description of the lexicon, and is specified for each language depending on the specificities of it.

A complete description of the model for morphology is available in GENELEX reports (public domain) on the morphological layer, available both in English and in French. [6]

There are different kind of MUs:

1. *Autonomous morphological units*:

    - Simple morphological units (usual entries of dictionaries) (UM_S)
    - Agglutinated morphological units (for instance the agglutination of a preposition and a determiner in French, Spanish and Italian)
    - Compound morphological units (UM_C) (continuous frozen compound words-vs. idioms or compound units that can be described at the syntactic level) have their forms calculated from the forms of their components and special separators between components if necessary.

2. *Non-autonomous morphological units*:

---

[6]GENELEX Consortium Report on the Morphological Layer V 3.3, November 2, 1994

- Affixes (prefixes, suffixes, infixes) (UM_Aff) as elements for derivation and for predictions on neologism.
- Units that can be found only in compound words.(non-autonomous UM_S)

MU have different characteristics:

1. Morphosyntactic category (or Part of speech) and sometimes sub-category

The list of values for these attributes is an instantiation, for each language, of EAGLES recommendations for morphosyntactic information in the lexicon

At least one and possibly several written forms (Graphical Morphological Unit or UMG) are associated to simple units (UM_S) Mention may be made of their stem(s); variants can be expressed through several UM associated to the same UM_S.

2. Inflected forms:

- Morphological features:

A UM_S has relevant combinations of morphological features that it can bear and is a characteristic of its paradigm. The list of these morphological features (Gender, Number, Mood, Person, for instance.) as well as the possible values are to be conformant to the EAGLES morphosyntax recommendations instantiated for each language.

- Inflectional behaviour of simple words

It may be described through two alternative explicit descriptions of methods of computation :

Addition of an affix to a stem .

Removal or addition of characters to a base form for a written morphological unit (which relieves the necessity for a morphemic description of UMGs).

Each "formula" to calculate an inflected form is associated to a set of morphological features. MFGs (inflectional Paradigms) are sets of associations of "formulas" to combination of morphological features ; they are associated to UMGs, and shared by different UM_S having the same inflectional properties.

- Inflectional behaviour of compound words

The inflectional system of compounds morphological units describes the inflections of the compound in relation with the variations (inflections) of each of its component.

3. Derivation:

Derivation relation are expressed through an ordered set of links oriented from the derived element to its internal components, that can be characterized according to status (base, suffix, etc.)

Affixes have derivational characteristics encoded : the selected morphosyntactic category, the result category and possibly its inflection mode.

4. Abridged forms:

These are expressed through "short form" relations between UMs, that may be typed according to the nature of the abbreviating mechanism (acronym, use of initials, abbreviation, etc.)

5. Usage values

UM or UMG can bear combinations of usage values (rare, archaic, colloquial, etc.) and geographic particularities (British english/American english), as well as frequency and dating .

The model is instantiated for each language, giving a list of features and their possible values: list of morphosyntactic categories and subcategories, list of morphological features and values associated to the descriptions of inflection mode of those categories/subcategories, list of types of affixes, of abridged relation.

3.3.8.1.2 Syntax

The syntactic level of the model deals coherently with all categories in a same descriptive language which thus allows to express an instantiation of EAGLES recommendations of syntactic description of verbs coherently inserted in the global architecture of the lexicon. It deals with the syntactic description for all categories for simple units as well as for non-frozen compounds (called syntactic compounds).

GENELEX syntactic description gives possibility to very fine-grained description. Some extension has been added to deal with some aspects of EAGLES syntactic recommendations for verbs when it was necessary, for instance as in the case of the PAROLE project, which represents an important instantiation of GENELEX.

A complete description of the GENELEX model for syntax is available in GENELEX reports on the syntactical layer (GENELEX, 93), public domain, available both in English and in French.The syntactic level is where all the information about the lexical unit syntactic behaviour is described, and especially what is not predictable by just knowing its morphosyntactic category and subcategory (which is already a very rough classification for the kind of syntactic behaviour to be expected, borne by the UM). As for morphology, complex and structured objects are defined in order to support explicitly the syntactic properties of each lexicon unit.

The GENELEX model syntactic level ìdeals with

- subcategorization, including functions of subcategorized complements and possibly thematic roles and semantic classes
- characteristics of the lexical unit when associated to a subcategorization frame

- control
- diathesis alternations
- pronominalization
- linear order constraints
- constraints on the syntactic context where the lexical unit is inserted (as subcategorized or modifier) associated eventually to its subcategorizing properties (necessary mainly for non-verb elements)
- syntactic compounds (idioms)

The GENELEX model for syntax can be described as follows :

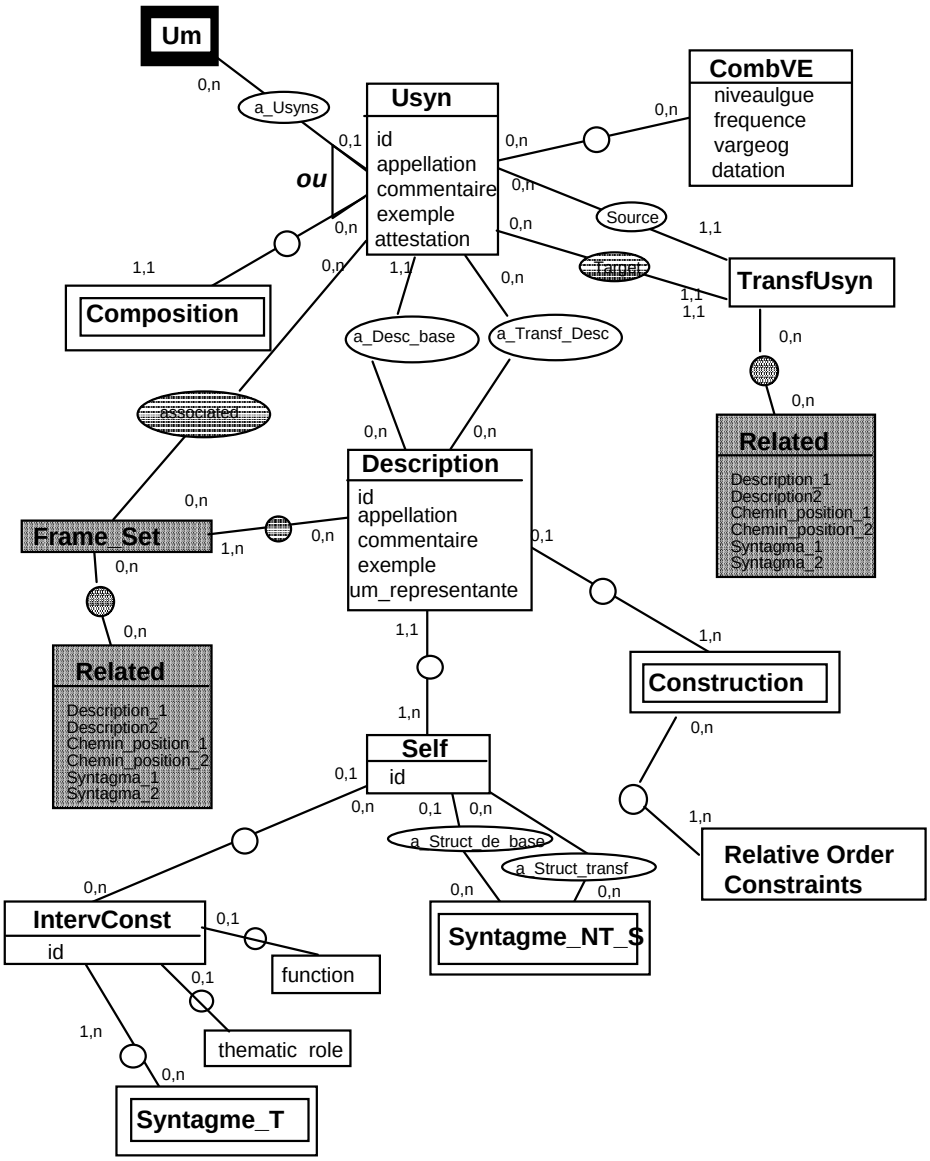


Fig. 18: Te GENELEX modelex for syntax

Mapping GENELEX-EAGLES objects

Description <--> Frame

Construction <--> List of Slots

Position_C <--> Slot

Syntagme(_T or _NT_C) <--> Slot Realization

and

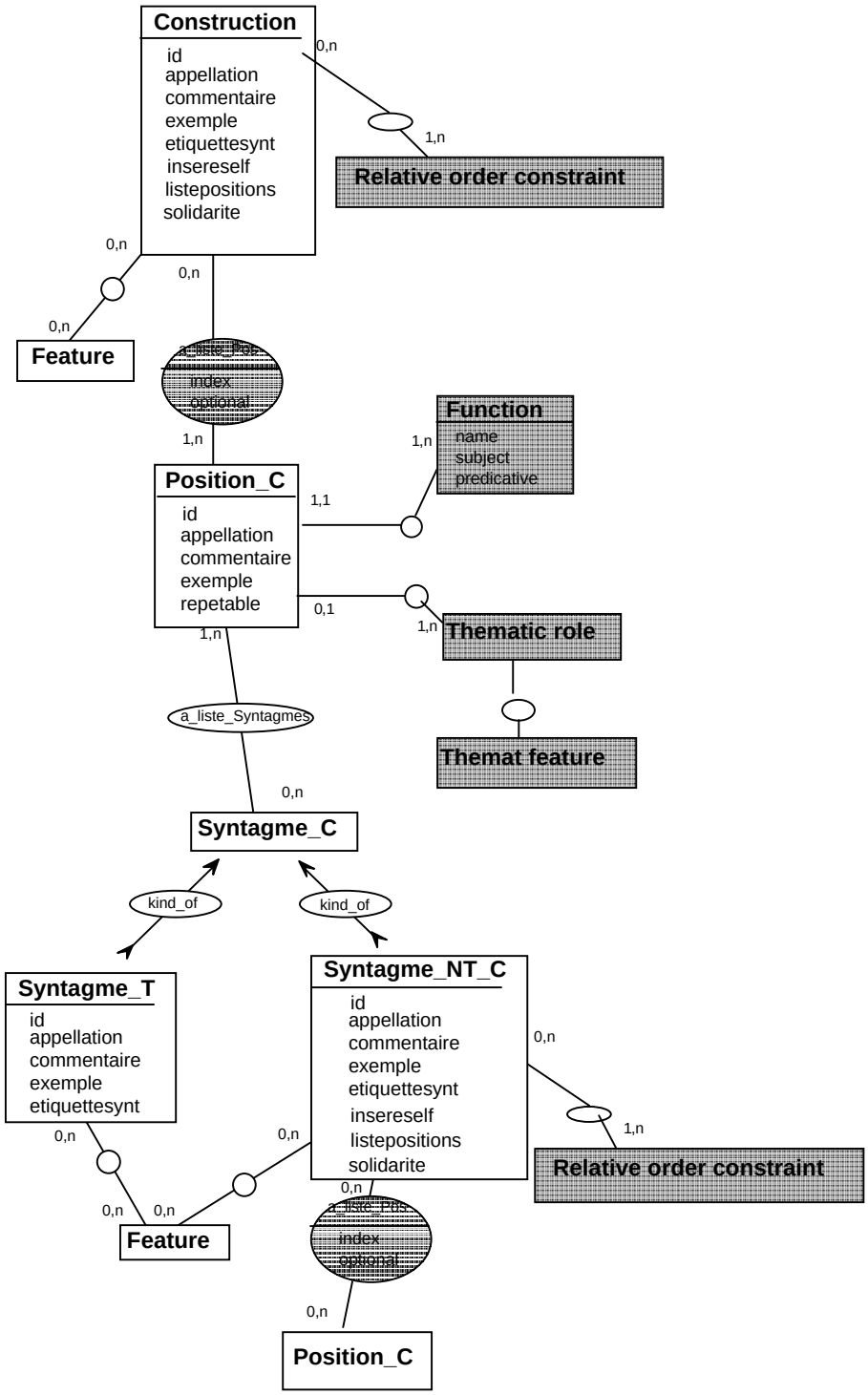


Fig. 19

Mapping GENELEX-EAGLES objects

Construction <--> List of Slots

Position_C <--> Slot

Syntagme(_T or _NT_C) <--> Slot Realization

The original core of the GENELEX model has been extended to reach a full compatibility with the EAGLES recommendations. These extensions - marked in grey - the following :

- the explicitation of FRAME_SET as a separate object. Frame sets are only implicitely present in the GENELEX model, as well as Related object
- the relative order constraint object, which doesn't exist in GENELEX model
- the fact that Function and Thematic role are attributed objects and not only attributes
- the index and optionality attributes on the relation between Construction or Syntagme_NT_C and Position

The mapping from this model to EAGLES is pretty straight.

Frame_Set (when explicitely present)<--> Frame_set
Set of Descriptions of Usyns linked to the

same UM with the Related of the different

TransfUsyns linking those Usyn when implicit<--> Frame_set

Related	<-->	Related
Description	<-->	Frame
Self	<-->	Self
Function	<-->	Function
Thematic role	<-->	Thematic role
Position_C	<-->	Slot
Syntagme_NT_C and Syntagme_T	<-->	Slot Realization

Etiquettesynt + the set of non-semantic features
associated to syntagma + Function on Position <--> Category

Special controlled_by feature	<-->	Controlled_by
Special coref feature	<-->	Obviates
Control feature (at construction level)	<-->	Control
Listepositions information	<--->	Pre/Post information
Optional	<-->	Optionality
Relative_order constraints	<-->	Rel-order
Non-semantic features	<-->	Features
Semantic class feature	<-->	Semantic class

Semantic class Feature on Syntagma
+ Thematic role on Position <--> Semantics

A Syntactic Unit (SyntU) can be simple or compound. A simple SyntU is one syntactic behaviour of one and only one MU, and this behaviour is specified by different descriptive elements. These are complex and structured objects. SyntUs are associated to one or more Semantic Units (SemUs), each Semantic Unit represents one acception of the lexical entry and is associated to syntactic contexts via the SyntUs. The mapping between syntax and semantics is explicitely described by different descriptive elements used to filter the syntactic structure (reject some Syntagmas, add some more constraints on Syntagmas, ...), to semantically constrain or enrich the semantic interpretation of some Position-Slot, to link semantic Arguments to syntactic Positions.

A SyntU (Usyn) is characterized by at least one Base Description (Frame), and possibly several other Descriptions which are surface alternations of the base Descriptions The Base Frame is chosen as reference one for explicitly describing the correspondance between syntax and semantics. Anyway, the Frames associated to one SyntU are describing different syntactic context that are closely related surface alternations associated to the same set of meaning(s). These "transformed" descriptions are supposed to be very close to the Base Description: for instance, passive descriptions, some impersonnal descriptions. These closely linked alternative frames are a possibility of the model to avoid splitting Usynts, but it is not supposed to deal with all the usual "alternation" phenomena wich are dealt with different Usyns associated to different Descriptions and linked by a TransfUsyn descriptive element that explicitely relates two Descriptions, which is the regular way to deal with alternation. It is important to notice that two Usyns can be associated to the same meaning (by two different correspondence set of information).

A SyntU is a "private" object as it is associated to one UM if it is simple. Set of frames (if used in a lexicon) and frames are usually shared by many different lexical elements, as they represent syntactic properties of lexical items.

Frame_Set is a set of Description-Frame that are related through the Related objects that links Positions or Syntagmas. The GENELEX model lets open the criteria to group Descriptions in a Frame_Set, but it is a good descriptive object to capture some generalization on a set of regular alternation and to represent something like "deep-syntax structure".

These descriptive objects of the syntactic level consist mostly of:

* Self (the lexical unit characteristics or constraints)
* Description (Frame) (a syntactic behaviour as the association of one Self and a Construction)
* Construction : list of Positions (slots) (a complementation frame inserted or not in a wider context)
* Position (Slot) (a complement or an element of context)
* Syntagma (Slot realization) (one of the possible surface realization of a slot)
* Typed-feature (restriction to be added on Slot realizations or on Self)
* Frame Set (a set of possibly alternated Frame)

SyntUs can be linked by alternation relation. It deals with encoding of alternation without taking into account the fact that the meaning is different or not. It also allows to encode syntactic derivation linking two SyntUs associated to two differnt UM, verbs and deverbal nouns for instance.

Some notions are parametrizable, specially the key notion of Position (Slot) whose definition may be purely syntactic (distribution paradigm, function) or syntactico-semantic (not only distribution paradigm, function but also theta-roles, semantic classes).

In the following subsections, we illustrate some syntactic phenomena that allow for a representation in the GENELEX format.

3.3.8.1.2.1 Subcategorization

For verbs and for the global characterization of main categories, subcategorization is the main point to describe. It is described by at least one Base Description or Frame associated to one Usyn A Description-Frame, is a complement pattern that can include the subject of verbs. Complementation patterns are not defined a priori, but instead the lexicographer when specifying the encoding of properties has at his disposal basic objects (Position-Slot, Syntagma-Slot realization, features) whose assembling produces a posteriori a finite set of patterns. Complement patterns Frame can be defined for verbs, but also for nouns, adjectives, adverbs and even other categories if necessary.

The model does not choose between fine grained or rough information. Again the assembling of basic objects provides the means of recording fined grained or rough information.

The number of complements corresponds to the number of Position-Slots directly surrounding the lexical entry in the Description -Frame. There is no predefined limit to the number of complements. The maximum number is determined by the lexicographer when applying the criteria to identify the elements of the Description-Frame, i.e subcategorized Position-Slots, that can be essential complements or special modifiers (obligatory or constrained by the lexical entry). Optional complements are Positions-Slots as well as obligatory ones.

Each Position-Slot may be defined as obligatory or optional.

There is an encoding criteria to determine when it is necessary to consider that there are two different Descriptions-Frames (maybe related in the same set of frames) and not one Description-Frame with optional Positions: all the combinations of realizations of Positions-Slots in a Description must be possible surface realizations.

Positions-Slots (or complements when subcategorized) may be realized either by a terminal or by a non-terminal category, and this category must be specified:

- terminal: the same categories as determined at morphological level: noun, adjective, adverb, verb, preposition, conjunction, interjection, determiner, pronoun, particle...

- non terminal: NP, PP, AP, ADVP, DETP, VP, S.(the phrases associated to main categories as heads).

Typed-features can be added on to the category to provide with more fine grained restrictions. Lexical, morphological, morpho-syntactic, syntactic, syntactico-semantic and semantic features are avalaible.

Phrases setting up alternatives for the realization of a same complement are gathered into a distribution paradigm. The Pronouns to be used can be specified there in the distribution of the Position.

When it is necessary to constrain structurally a Syntagma-Slot Realization, it is possible to rewrite it (partially or completely) as a list of Position-Slots

A Position is characterized by its distribution and its function, and for some approach to syntax with some basic semantics, eventually thematic role or semantic class restriction on its realizations.

This way a single Position can be encoded for e.g. objective complements and their different realizations: NP, that-clause, infinitival.

Lexical selection in complementation patterns includes :

bound prepositions: to/for/with/on/...

complementizer: that/whether/...

impersonal subjects: it/there

clitics: he/him

Characteristics of the lexical unit when associated to the frame: a special object called Self carries this information, expressed by means of the features that are used to constrain the Syntagmas-realizations of Slots , and some special ones as, for instance, the one for verbs that, depending on the languages, expresses the auxiliary to form compound tenses or to passivize. To deal with the rewriting possibilty and with categories that are syntactically described by specifying their insertion context or the context where they occur as subcategorized elements, Self can bear Function and Thematic role. It allows to describe syntagmas where Self (the lexical unit when inserted in the context) is not the syntactic head, and where all functions of Position-Slots are not specified in relation to the Self Unit, but in relation to another element which bears the function Head.

3.3.8.1.2.2 Alternations

Diathesis alternations such as ergative/inchoative and possibly active/passive alternations, is handled: in a descriptive way, by a link relating two Descriptions-Frames and their Positions-Slots, (and posibly specifying syntagmas) ; Both linked Descriptions are associated to either two existing Usyns or to the same Usyns for some special very close alternation links preserving the meaning. Those Descriptions can be explicitely linked to a Frame_set or not.

Some other alternations can be described by the alternation relation linking two diferent SyntUs associated to two different UM and the Positions-Slots of their base Description-Frame. This allows to link a deverbal nominal Usyn to the verbal Usyn it is "syntactically derived".

3.3.8.1.2.3 Linear order constraints

Linear order constraints can be expressed when the free ordering of syntagmas, which is described by the grammar and not by the lexicon, is more constrained for a special lexical entry. Order relations between two Positions or two Syntagmas or relative to Self can be borne by the Frame. A special attribute allows to give the status of the constraint: preferential or mandatory.

3.3.8.1.2.4 Insertion context

Constraints on the syntactic context where the lexical units (together with its complements) is inserted can be expressed. This concerns mainly adjectives or adverbs, and some nouns. It permits for instance to describe the behaviour of an adjective as left or right attribute by describing the prototypical NP structure where it is inserted; it permits to describe that some adjective enters in sentences with impersonal subject, and so on.

These constraints on the insertion context are expressed through the possibility of describing the context as a tree by the rewriting of Positions-Slots as associated to a list of Positions: Positions-Slots can be described as a complex structure, and associated to the list of Positions-Slots that is their rewriting. So, with this powerful mechanism, a Description-Frame can coherently express both the insertion context and the subcategorized context associated to the SyntU in a same tree..

3.3.8.1.2.5 Syntactic compounds (idioms)

Some compounds are not a continuous sequence of their components and support many variations in surface, even though they have a meaning as a whole. Those idioms are described as Compound Syntactic Units. As external behaviour, they are described as simple units, by means of Descriptions and Frame_sets , but their Self bears an additional information: the information about their internal structure (Syntagme_NT_S) and the SyntU has the information of the list of components (Composition); the internal structures can be shared, independently of the lexicalisation of the leaves of the structure, dealt with by pointer to Composition elements.. Alternatives of lexicalisation can be dealt with in one compound SyntU, and also the possible alternations for the internal structure, and the interactions between the external structure (Construction : list of Positions-Slots) and the internal one (Syntagme_NT_S: a list of Positions-Slots), and how they can (or they cannot) alternate in the surface.

3.3.8.1.3 Semantics

The GENELEX semantic layer, and is a compatible extension of the two preceeding layers. We give now just a quick overview of the semantic layer. For more details, the GENELEX Report on the Semantic Layer is the reference (GENELEX, 1994b). An important instantiation of the GENELEX semantic layer is represented by the SIMPLE model, illustrated in 3.2.5.3.

The mapping between syntactic and semantics level is explicitely described, giving possibility to filter out some syntactic realizations on syntactic or also semantic criteria.

It deals with two distinct sublevels of representation: the first sublevel may be seen as strictly the domain of lexical semantics, whereas the second one aims at representing a more cognitive type of content.

The main entities for lexical semantics are **Semantic Units (Usem**). They are closely connected with the syntactic level, as every USem has to be related to at least one Usyn. This relation may be restricted through constraints on the Usyn itself and/or by filters on the positions it governs in a Frame ; the relation may also precisely state the way semantic arguments of a predicate match syntactic positions governed by a USyn, and give default semantic values for implicit arguments (for example when an absolute construction is possible).

A USem may be connected to a linguistic predicate, which can summarise semantic information about predicative USem s. This connection may take various modes: a predicate may be lexicalized by one or more USem s, and so may its arguments. For example, USem 1 can be one of the lexicalizations of predicate P; USem 2 may be the typical lexicalization of P incorporating argument 1, etc.

USems can be described by means of:

1. Semantic features: Analytic description is carried out through a set of values of semantic features that range from classical componential features to pragmatic features and including information on domain, connotative value, etc.

2. Cross-references: These can be expressed through a set of specialized relationships:

* Paradigmatic relationships, such as hypo/hyperonymy, synonymy, meronymy, etc.;

* Semantic derivation relationships, that may represent derivation according to the meaning (with or without morphological motivation), or information related to typicality

(e.g.\ USem1 is the typical location for USem 2 activity; USem 3 is the typical instrument for USem 4 action, etc.), when not expressed through a predicate;

* Collocation preference relationships (e.g.\ USem 1 is the preferred intensifier for USem 2 ; USem 3 is the support verb for some aspect of USem 4 action, etc.).

3. Predicates (that are one of the basic descriptive elements) may also be described in these two ways, although the features and relationships available for their description are less varied.

4. Cognitive generalizations. The semantic level also offers the possibility of abstracting cognitive units (**concepts**) from USem sand/or from predicates. Such units may be useful in factoring information about a lexical equivalence class (for instance synonyms carrying different connotative values), and also in establishing content units that do not have a lexical realisation for a given language: such lexical gaps may need to be filled for the purpose of establishing taxonomies or for representing terminological data.

3.3.8.1.4 Multilingual links

Finally, the structure of the syntactic and semantic level paves the way for establishing multilingual links as defined in model for multilinguality, and it is a compatible extension of the

two preceeding layers. We give now just a quick overview of the multilingual layer. For more details, the GENELEX Report on the Multilingualism is the reference (GENELEX, 1994c).

The multilinguality can be dealt with in two complementary approaches.

1. **Contrastive approach.** A contrastive description of multilingual correspondance, set at the level of lexical semantics, establishes multilingual correspondance between lexical semantics elements (SEMU or Predicate) of one language with lexical semantics elements in another language. Syntactic contexts of each languages are always implicitly linked (through the monolingual correspondance between syntax and semantics). Multilingual links are thus mainly at the level of semantics. Anyway, Syntactic Units can be explicitly multilingually linked to preferencially associate one particular syntactic realization of one meaning in one language to another particular syntactic realization of one meaning in another language, depending on the approach to multilingualism and on the kind of linguistic object to be linked. Some filters can apply (on syntax and semantics) when establishing multilingual links. They are expressed in the same language as filtering from semantic to syntax, i.e. the same descriptive objects.

2. **Interlingua approach**. An interlingual description based upon sharing of so called "primitives" (concepts, predicates, and features) can be made, and, even if this prospect seems rather fuzzy at present for direct NLP applications, such a possibility could turn out to be very useful for terminological purposes. This complementary approach is a good way to give possibility to different fine grained description; interlingua approach neglects very subtle and language dependant meaning distinctions

3.3.8.2 Data representation

The GENELEX model for lexicon is expressed in an SGML DTD, which makes explicit the different descriptive elements to be used within the lexical description, their relations and the global structure of the whole as well as the details of the features and possible values, and the optionality or mandatoriness of the information.

Some constraints are not expressed within SGML formalism, so they will be expressed in natural language within the commented SGML DTD and then translated in the formal language to Integrity constraints verifications by the software.

3.3.8.3 Extensions to other information

The GENELEX model as is at the moment doesn't deal at all with spoken aspects of the lexicon nor terminological ones. However, the global architecture of the lexicon is designed to easily be extended as to represent or to be connected to that kind of information.

As in GENELEX model, phonetical information can be added at the level of the Morphological Unit, for instance as Phonetical Unit associated to the UM as is the Graphical Unit. That

connection to the "spoken characteristics" of the lexical entry permits to complete the lexicon, and to make that spoken resources might have access to information on the syntax and semantics of the entry.

A similar possibility exists for terminology to be connected to a general lexicon. The connection is currently being defined in the project Transterm; the model of this connection of terminological data to general lexicon is defined in an approach compatible with the GENELEX model (especially the semantic layer). So, the model (and the encoded data) could be easily and coherently extended, to make the generic lexical database (designed at first for written general resources) deal in a generic way with other application needs.

Table 17. Synoptic table of information types in GENELEX

	Entry component		Present	Representation in Genelex
1	Headword		✓	It is the value of the `id` attribute in the Morphological unit
2	phonetic transcription			
3	variant form			
4	inflected form		✓	Morphological units contain a link to the inflectional tables where number, gender, mood, tense information is contained, as well as the particular way in which the lexeme is inflected
5	cross-reference			
6	**Morphosyntactic information**			
	a	Part-of-speech marker	✓	Value of the gramcat attribute in the Morphological unit
	b	Inflectional class	✓	Morphological units contain a link to the inflectional tables where number, gender, mood, tense information is contained, as well as the particular forms of a given entry
	c	Derivation	✓	Cross part of speech relations are marked through derivational semantic relations between SemUs
	d	Gender	✓	Expressed in the Ginp associated to a Morphological Unit
	e	Number	✓	Expressed in the Ginp associated to a Morphological Unit
	f	Mass vs. Count	✓	Expressed in the Morphological Unit
	g	Gradation	✓	Expressed in the Morphological Unit
7	subdivision counter			
8	entry subdivision		✓	Value of the attribute id in the `SemU` object
9	sense indicator		✓	This information is captured by the values of the attributes `naming`, `example` and `comment`, which conjointly give clues to show the specific sense encoded in the `SemU`
10	linguistic label		✓	Only for information about the terminological domain
11	**Syntactic information**			
	a	Subcategorization frame	✓	Described in the Syntactic Units specifying the number of positions, the syntactic realization (type of phrase, introducer, etc.). Each syntactic description is then linked to a Semantic Units, and the arguments structures are linked to their syntactic realizations
	b	Obligatority of complements	✓	Marked in the Syntactic Unit
	c	Auxiliary	✓	Marked in the Self object associated to a Syntactic Unit
	d	Light or support verb construction	✓	
	e	Periphrastic constructions	✓	

	f	Phrasal verbs	✓	
	g	Collocator	✓	Optionally encoded in the semantic layer: typical subject, typical object, etc.
	h	Alternations	✓	Represented in terms of syntactic descriptions (i.e. subcategorization structures) linked in a Frameset
12		**Semantic information**		
	a	Semantic Type	✓	Represented as link between a Semantic Unit and a node in the Ontology of semantic types
	b	Argument Structure	✓	Represented in the Predicative Representation associated to Semantic Units: it contains a link between the Semantic Unit and a predicate, on turn defined in terms of the number of arguments, their thematic roles, and selectional preferences
	c	Semantic relations	✓	Represented as relations between Semantic Units (e.g. hyperonymy, meronymy, and many others)
	d	Regular polysemy	✓	Represented as relations between Semantic Units
	e	Domain	✓	Represented as link between a Semantic Unit and a node in a hierarchy of domains
	f	Decomposition	✓	Represented as relations among predicates
13	translation		✓	
14	gloss		✓	In the attribute `freedefinition` a gloss is specified, as derived from a medium-sized monolingual dictionary
15	Near-equivalent			
16	Example phrase (straightforward)		✓	This is the value of the attribute `example`
17	Example phrase (problematic)			
18	multiword unit		✓	Represented as compund syntactic units
19	subheadword (secondary headword)			
20	usage note		✓	
21	frequency			

4 Synoptic Grids

In what follows, to ease comparison among different surveyed resources, we give an overview about how the information is distributed in the resources, maintaining the subdivision in three different types of resources: MRDs, Computational Lexicons and Resources for MT systems.

.

4.1 MRDs

Table 18: Synoptic Grid of the information types in MRDs

	Entry component		Information content	Collins	Gem	Oxford Hachette	Oxford	Van Dale
1	headword		lexical form(s) of the headword: how the headword is spelt	✓	✓	✓	✓	✓
2	Phonetic transcription		how the headword (or variant form etc.) is pronounced (in *International Phonetic Alphabet*)	✓		✓	✓	
3	variant form		alternative spelling of headword or slight variation in the form of this word	✓	✓	✓	✓	✓
4	inflected form		other grammatical forms of the lemma (headword)	✓	✓	✓	✓	
5	Cross-reference		indication of another headword whose entry holds relevant information, or some other part of the dictionary where this may be found	✓	✓	✓	✓	
6	**Morphosyntactic information**							
	a	Part-of-speech marker	part of speech of the headword (or the secondary headword)	✓	✓	✓	✓	✓
	b	Inflectional class	Inflectional paradigm of the entry					
	c	Derivation	Cross-part-of-speech-information, morphologically derived forms					
	d	Gender	Information about the gender of the entry in SL and TL	✓	✓	✓	✓	✓
	e	Number	Information about the grammatical number of the entry in SL and TL	✓	✓	✓	✓	
	f	Mass vs. Count	Information whether a noun is mass or count, in SL and TL					

				Collins	Gem	Oxford-Hachett	Oxford	Van Dale
	g	Gradation	For adverbs and adjectives	✓	✓	✓	✓	✓
7	Subdivision counter		indicates the start of new section or subsection (*'sense'*)	✓	✓	✓	✓	✓
8	Entry subdivision		separate section or subsection in entry (often called *dictionary sense*)	✓	✓	✓	✓	✓
9	Sense indicator		synonym or paraphrase of headword in this sense, or other brief sense clue indicating specific sense of SL or TL item	✓	✓	✓	✓	✓
10	linguistic label		the style, register, regional variety, etc. of the SL or TL item	✓	✓	✓	✓	✓
11	**Syntactic Information**							
	a	Subcategorization frame	(i.) Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.	✓	✓	✓	✓	
	b	Obligatority of complements	Information whether a certain complement is obligatory or not					

				Collins	Gem	Oxford-Hachette	Oxford	Van Dale
	c	Auxiliary	Which type of auxiliary is selected by a given predicate (in certain languages auxiliary selection is related to issues like unaccusativity, which on turn lies at the interface between lexicon and syntax)					
	d	Light or support verb construction	Constructions with light verbs	✓	✓	✓	✓	
	e	Periphrastic constructions	Constructions containing periphrasis, usage, semantic value, etc.	✓	✓	✓	✓	
	f	Phrasal verbs	Particular representation of phrasal constructions	✓	✓	✓	✓	✓
	g	Collocator	(i.) typical subject /object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented etc.	✓	✓	✓	✓	✓
	h	Alternations	Syntactic alternations an entry can enter into					
12	**Semantic Information**							
	a	Semantic type	Reference to an ontology of types which are used to classify word senses		✓			
	b	Argument structure	Argument frames, plus semantic information identifying the type of the arguments, selectional constraints, etc.		✓			

				Collins	Gem	Oxford-Hachette	Oxford	Van Dale
	C	Semantic relations	Different types of relations (e.g. synonymy, antonymy, meronymy, hyperonymy, Qualia Roles, etc.) between word senses, etc.	✓	✓			
	d	Regular polysemy	Representation of regular polysemous alternations					
	e	Domain	Information concerning the terminological domain to which a given sense belongs	✓	✓	✓	✓	✓
	f	Decomposition	Representation of relevant meaning component, e.g. causativity, agentivity, motion, etc.					
13	Translation		TL equivalent of SL item	✓	✓	✓	✓	✓
14	Gloss		TL explanation of meaning of an SL item which has no direct equivalent in the TL	✓	✓	✓	✓	✓
15	Near-equivalent		TL item corresponding to an SL item which has no direct equivalent in the TL	✓	✓	✓	✓	✓
16	Example phrase (straightforward)		a phrase or sentence illustrating the non-idiomatic use of the headword, in a context where the TL equivalent is virtually a word-to-word translation	✓	✓	✓	✓	✓

			Collins	Gem	Oxford-Hachette	Oxford	Van Dale
17	Example phrase (problematic)	a phrase or sentence illustrating a non-idiomatic use of headword in a context where a specific TL equivalent is required (*i.e. an SL example which is easily understandable for the TL speaker, but presents translation problems for the SL speaker*)	✓	✓	✓	✓	✓
18	multiword unit	(idiomatic) multiword expression (MWE) containing the headword *(the term MWE covers idioms, fixed & semi-fixed collocations, compounds etc*.)	✓	✓	✓	✓	✓
19	Subheadword *also* secondary headword	lemma morphologically related to the headword, figuring as head of a sub-entry (*subheadwords can be compounds, phrasal verbs, etc.)*		✓	✓	✓	
20	usage note	how the headword is used; 'macro' information which cannot appear at every appropriate entry; warning of cultural differences between the two languages; etc.	✓	✓	✓		
21	Frequency	Information about the frequency of the entry				✓	

4.2 Computational Lexicons

Table 19: Synoptic Grid of the information types in Computational Lexicons

	Entry component		Information content	Collins-Robert	FrameNet	Euro(/Ital) WordNet	PAROLE-Simple
1	headword		lexical form(s) of the headword: how the headword is spelt	✓	✓	✓	✓
2	Phonetic transcription		how the headword (or variant form etc.) is pronounced (in *International Phonetic Alphabet*)				
3	variant form		alternative spelling of headword or slight variation in the form of this word		✓	✓	
4	inflected form		other grammatical forms of the lemma (headword)	✓			✓
5	Cross-reference		indication of another headword whose entry holds relevant information, or some other part of the dictionary where this may be found				
6	**Morphosyntactic Information**						
	a	Part-of-speech marker	part of speech of the headword (or the secondary headword)	✓	✓	✓	✓
	b	Inflectional class	Inflectional paradigm of the entry				✓
	c	Derivation	Cross-part-of-speech-information, morphologically derived forms	✓		✓	✓
	d	Gender	Information about the gender of the entry in SL and TL	✓		✓	✓
	e	Number	Information about the grammatical number of the entry in SL and TL	✓		✓	✓

				Collins-Robert	FrameNet	Euro(Ital) WordNet	PAROLE-SIMPLE
	f	Mass vs. Count	Information whether a noun is mass or count, in SL and TL				✓
	g	Gradation	For adverbs and adjectives			✓	✓
7	Subdivision counter		indicates the start of new section or subsection (*'sense'*)		✓		
8	Entry subdivision		separate section or subsection in entry (often called *dictionary sense*)			✓	✓
9	Sense indicator		synonym or paraphrase of headword in this sense, or other brief sense clue indicating specific sense of SL or TL item	✓	✓	✓	✓
10	linguistic label		the style, register, regional variety, etc. of the SL or TL item	✓		✓	✓
11	**Syntactic Information**						
	a	Subcategorization frame	(i.)Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.		✓		✓
	b	Obligatority of complements	Information whether a certain complement is obligatory or not				✓

				Collins-Robert	FrameNet	Euro(Ital) WordNet	PAROLE-Simple
	c	Auxiliary	Which type of auxiliary is selected by a given predicate (in certain languages auxiliary selection is related to issues like unaccusativity, which on turn lies at the interface between lexicon and syntax)				✓
	d	Light or support verb construction	Constructions with light verbs	✓	✓		
	e	Periphrastic constructions	Constructions containing periphrasis, usage, semantic value, etc.				
	f	Phrasal verbs	Particular representation of phrasal constructions	✓	✓		✓
	g	Collocator	(i.) typical subject /object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented) etc.	✓			✓
	h	Alternations	Syntactic alternations an entry can enter into		✓		✓
12	**Semantic Information**						
	a	Semantic type	Reference to an ontology of types which are used to classify word senses			✓	✓
	b	Argument structure	Argument frames, plus semantic information identifying the type of the arguments, selectional constraints, etc.		✓	✓	✓

				Collins-Robert	**FrameNet**	**Euro(Ital) WordNet**	**PAROLE-Simple**
	c	Semantic relations	Different types of relations (e.g. synonymy, antonymy, meronymy, hyperonymy, Qualia Roles, etc.) between word senses, etc.	✓	✓	✓	✓
	d	Regular polysemy	Representation of regular polysemous alternations			✓	✓
	e	Domain	Information concerning the terminological domain to which a given sense belongs	✓	? ✓	✓	✓
	f	Decomposition	Representation of relevant meaning component, e.g. causativity, agentivity, motion, etc.				
13	Translation		TL equivalent of SL item	✓	✓	✓	
14	Gloss		TL explanation of meaning of an SL item which has no direct equivalent in the TL	✓		✓	✓
15	Near-equivalent		TL item corresponding to an SL item which has no direct equivalent in the TL			✓	
16	Example phrase (straightforward)		a phrase or sentence illustrating the non-idiomatic use of the headword, in a context where the TL equivalent is virtually a word-to-word translation		✓	✓	✓

			Collins-Robert	FrameNet	Euro(Ital) WordNet	PAROLE-Simple
17	Example phrase (problematic)	a phrase or sentence illustrating a non-idiomatic use of headword in a context where a specific TL equivalent is required (*i.e. an SL example which is easily understandable for the TL speaker, but presents translation problems for the SL speaker*)	✓		✓	
18	multiword unit	(idiomatic) multiword expression (MWE) containing the headword *(the term MWE covers idioms, fixed & semi-fixed collocations, compounds etc*.)	✓		✓	
19	Subheadword *also* secondary headword	lemma morphologically related to the headword, figuring as head of a sub-entry (*subheadwords can be compounds, phrasal verbs, etc.)*				
20	usage note	how the headword is used; 'macro' information which cannot appear at every appropriate entry; warning of cultural differences between the two languages; etc.				
21	Frequency	Information about the frequency of the entry		✓		

4.3 Resources for MT systems

Table 20: Synoptic Grid of the information types in resources for MT systems

	Entry component		Information content	Eurotra	METAL	EDR	Microsoft	GENELEX
1	headword		lexical form(s) of the headword: how the headword is spelt	✓	✓	✓	✓	✓
2	Phonetic transcription		how the headword (or variant form etc.) is pronounced (in *International Phonetic Alphabet*)			✓	✓	
3	variant form		alternative spelling of headword or slight variation in the form of this word	✓	✓	✓	✓	
4	inflected form		other grammatical forms of the lemma (headword)		✓	✓	✓	✓
5	Cross-reference		indication of another headword whose entry holds relevant information, or some other part of the dictionary where this may be found			✓	✓	
6	**Morphosyntactic Information**							
	a	Part-of-speech marker	part of speech of the headword (or the secondary headword)	✓	✓	✓	✓	✓
	b	Inflectional class	Inflectional paradigm of the entry	✓	✓	✓	✓	✓
	c	Derivation	Cross-part-of-speech-information, morphologically derived forms	✓		✓	✓	✓
	d	Gender	Information about the gender of the entry in SL and TL	✓	✓		✓	✓
	e	Number	Information about the grammatical number of the entry in SL and TL	✓	✓	✓	✓	✓

				EUROTRA	METAL	EDR	Microsoft	GENELEX
	F	Mass vs. Count	Information whether a noun is mass or count, in SL and TL	✓	✓	✓	✓	✓
	g	Gradation	For adverbs and adjectives		✓	✓	✓	✓
7	Subdivision counter		indicates the start of new section or subsection (*'sense'*)				✓	
8	Entry subdivision		separate section or subsection in entry (often called *dictionary sense*)				✓	✓
9	Sense indicator		synonym or paraphrase of headword in this sense, or other brief sense clue indicating specific sense of SL or TL item				✓	✓
10	linguistic label		the style, register, regional variety, etc. of the SL or TL item		✓	✓	✓	✓
11	**Syntactic Information**							
	a	Subcategorization frame	(i.)Number and types of complements (ii.) syntactic introducer of a complement (e.g. preposition, case, etc.) (iii.) type of syntactic representation (e.g. constituents, functional, etc.) etc.	✓	✓	✓	✓	✓
	b	Obligatority of complements	Information whether a certain complement is obligatory or not	✓	✓	✓		✓
	c	Auxiliary	Which type of auxiliary is selected by a given predicate (in certain languages auxiliary selection is related to issues like unaccusativity, which on turn lies at the interface between lexicon and syntax)	✓	✓	✓	✓	✓

				EUROTRA	METAL	EDR	Microsoft	GENELEX
	d	Light or support verb constructio n	Constructions with light verbs	✓		✓		✓
	e	Periphrastic constructio ns	Constructions containing periphrasis, usage, semantic value, etc.			✓		✓
	f	Phrasal verbs	Particular representation of phrasal constructions	✓		✓	✓	✓
	g	Collocator	(i.) typical subject /object of verb, noun modified by adjective etc. (ii.) type of collocation relation represented) etc.			✓	✓	✓
	h	Alternations	Syntactic alternations an entry can enter into	✓		✓	✓	✓
12		**Semantic Information**						
	a	Semantic type	Reference to an ontology of types which are used to classify word senses	✓	✓	✓		✓
	b	Argument structure	Argument frames, plus semantic information identifying the type of the arguments, selectional constraints, etc.	✓		✓		✓
	c	Semantic relations	Different types of relations (e.g. synonymy, antonymy, meronymy, hyperonymy, Qualia Roles, etc.) between word senses, etc.			✓	✓	✓
	d	Regular polysemy	Representation of regular polysemous alternations			✓		✓
	e	Domain	Information concerning the terminological domain to which a given sense belongs		✓	✓	✓	✓
	f	Decomposit ion	Representation of relevant meaning component, e.g. causativity, agentivity, motion, etc.			✓		✓

			EUROTRA	METAL	EDR	Microsoft	GENELEX
13	Translation	TL equivalent of SL item	✓	✓	✓	✓	✓
14	Gloss	TL explanation of meaning of an SL item which has no direct equivalent in the TL			✓		✓
15	Near-equivalent	TL item corresponding to an SL item which has no direct equivalent in the TL			✓	✓	
16	Example phrase (straightforward)	a phrase or sentence illustrating the non-idiomatic use of the headword, in a context where the TL equivalent is virtually a word-to-word translation			✓	✓	✓
17	Example phrase (problematic)	a phrase or sentence illustrating a non-idiomatic use of headword in a context where a specific TL equivalent is required (*i.e. an SL example which is easily understandable for the TL speaker, but presents translation problems for the SL speaker*)			✓	✓	
18	multiword unit	(idiomatic) multiword expression (MWE) containing the headword *(the term MWE covers idioms, fixed & semi-fixed collocations, compounds etc*.)	✓	✓	✓	✓	✓
19	Subheadword *also* secondary headword	lemma morphologically related to the headword, figuring as head of a sub-entry (*subheadwords can be compounds, phrasal verbs, etc.)*		✓	✓	✓	
20	usage note	how the headword is used; 'macro' information which cannot appear at every appropriate entry; warning of cultural differences between the two languages; etc.			✓	✓	✓

21	Frequency	Information about the frequency of the entry			✓	✓	

5 Case Study: examples of cross-lingual linguistic phenomena

We selected a set of linguistic phenomena we consider worthy of study in order to perform the mapping from the SL to the TL in many multilingual applications, such as Cross-lingual Information Retrieval and Extraction, Machine Translation, multilingual analysis and generation.

The examples of these phenomena have been circulated among the partners, collecting the translations for the involved languages in order to assemble a representative set of mappings between languages which require more than simple word-to-word correspondences or non trivial word to word correspondences.

As a matter of fact, only in simpliest cases a multilingual lexicon simply has to replace a lexical item in the source language with a corresponding lexical item in the target language that conveys roughly the same meaning.

Many mappings are much more complex than this, and can require additional information. For instance, modifiers in one language may become matrix verbs in another, or vice versa. One language may use inflectional morphology to capture something that is better expressed with a separate lexical item in another language. A verb with one argument may require a corresponding verb with an additional prepositional phrase to convey the same meaning in another language. These varying methods of expression in different languages are often referred to under the umbrella of structural divergences, and represent special challenges for multilingual lexicons.

Since there are as many different formats for capturing this type of information as there are translation systems, all of which are worthy of examination, we gathered information about how many of the systems handle each specific phenomenon; the lexicon format has to allow for adequate contextual and structural information to be represented so that structural divergences can be recognized and dealt with accordingly.

In what follows we present a preliminary classification of these relevant lexical phenomena. In particular, we illustre with examples how different computational lexicons and systems represent and encode each of the lexical phenomena.

5.1.1 Examples of the problem of selecting a target language equivalent

5.1.1.1 Sense distinctions according to syntactic subcategorization frames[7]

a) The verbs **[know/saber]** in English, Spanish, and Italian may get different translations depending on the syntactic type of complement.

E: know

I: sapere (+Comp) - *John knows that Mary is ill* (*Gianni sa che Maria è malata*)

S: saber (+Comp)

E: know

I: conoscere (+NP) - *John knows Mary* (*Gianni conosce Maria*)

S: saber (+VPinf) - *saber nadar* / *leer* / *conducir*

E: can (+VPinf) - *can swim* / *read* / *drive*

b) **[bestehen]** in German has as English translation either "insist on" or "consist of", depending on the German preposition used.

G: bestehen + subj + p_obj (auf)

E: insist on

G: bestehen + subj + p_obj (in)

E: consist of

[7] In the examples below, we used the following abbreviations: C: Catalan, E: English, F: French, G: German, K: Korean, I: Italian, P: Portoghese, S: Spanish.

5.1.1.1.1 Sense distinctions according to syntactic frames in Collins Gem

word : know

translation 1 : savoir

translation 2 : connaître

Semantic constraint on translation 2 : domain (person, author, place)

5.1.1.1.2 Sense distinctions according to syntactic frames in PAROLE-Simple

Relevant Information in P-S: Syntactic Unit

Italian	(a.) *sapere* (to know something): "Gianni sa la matematica" (John knows maths) "Gianni sa che Maria è malata" (John knows that Mary is ill) (b.) *sapere* (to be able to do something) "Gianni sa nuotare" (John can swim)	
Analysis in P-S	**SemU**: *sapere* (a.) →	**SemU**: *know* **Synt. Construction**: pos1 = NP; pos2 = That_S / NP
	SemU: *sapere* (b.) →	**SemU**: *can* **Synt. Construction**: pos1 = NP; pos2 = Inf_V

5.1.1.1.3 Sense distinctions according to syntactic frames in SYSTRAN

know .if_object_is a noun_clause then translate IT "sapere"

know .if_object_is_a_noun+HUMAN then translate IT "conoscere"

saber .if_governs_inf "nadar" then "know how"

saber .if_governs_infinitive" then translate EN "can/be able", else translate EN "know"

(priorities can be assigned to assure ordering of rules, e.g. the examples for saber in the order shown here)

bestehen .if_prep_complement_is "auf"_and_its_object_is_abstract then translate EN "insist (on)"

5.1.1.1.4 Sense Distinction according to syntactic frames in Lexical Conceptual Structure Lexicon

[know/saber]

E: know (+Comp)	- John knows Mary to be ill	[E-1]
	- John knows that Mary is ill	[E-2]
(+PP)	- John knows (of/about) Mary	[E-3]
S: saber		[S-1,2,3]
E: know (+NP)	- John knows Mary	[E-4]
S: conocer		[S-4]

```
[E-1]:
-------------------------------------------------------------------
;; Grid: 29.5.a#1#_exp_perc_mod-prop(to)#
```

```
(DEFINE-WORD
  :DEF_WORD "know"
  :CLASS "29.5.a"
  :WN_SENSE (("1.5" 332083 333362 --) ("1.6" 400501 401762 411402))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT PERCEPTUAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * THING NIL NIL VAR 8)))
         (:MOD NIL POSITION AS CIRCUMSTANTIAL NIL 39
          ((:SUB NIL STATE *HEAD* NIL NIL 40)
           (:ARG * NIL NIL NIL VAR 28)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)) (28 (THING -) (CFORM INF) :OBLIGATORY))
  :COLLOCATIONS ((28 "to"))
)
----------------------------------------------------------------------

[E-2]:
----------------------------------------------------------------------
;; Grid: 29.5.b#1#_exp_prop(that)#

(DEFINE-WORD
  :DEF_WORD "know"
  :CLASS "29.5.b"
  :WN_SENSE (("1.5" 333362 333754) ("1.6" 401762 402210))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT CIRCUMSTANTIAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2) (:ARG * NIL NIL NIL VAR 27)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)) (27 (THING -) (CFORM FIN)))
  :COLLOCATIONS ((27 "that"))
)
----------------------------------------------------------------------

[E-3]:
----------------------------------------------------------------------
;; Grid: 29.5.c.i#1#_exp_perc(of,about)#

(DEFINE-WORD
  :DEF_WORD "know"
  :CLASS "29.5.c.i"
  :WN_SENSE (("1.5" 333362) ("1.6" 401762))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG * POSITION [ABOUT] PERCEPTUAL NIL 7
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG NIL THING NIL NIL VAR 8)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)))
)
----------------------------------------------------------------------

[E-4]:
----------------------------------------------------------------------
;; Grid: 29.5.c.ii#1#_exp_perc#

(DEFINE-WORD
```

```
  :DEF_WORD "know"
  :CLASS "29.5.c.ii"
  :WN_SENSE (("1.5" 333362) ("1.6" 401762))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION [ABOUT] PERCEPTUAL NIL 7
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * THING NIL NIL VAR 8)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)))
)
----------------------------------------------------------------------

[S-1]:
----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "saber"
  :GLOSS "know"
  :CLASS "29.5.a"
  :WN_SENSE (("1.5" 332083 333362 --) ("1.6" 400501 401762 411402))
  :LANGUAGE SPANISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT PERCEPTUAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * THING NIL NIL VAR 8)))
         (:MOD NIL POSITION AS CIRCUMSTANTIAL NIL 39
          ((:SUB NIL STATE *HEAD* NIL NIL 40)
           (:ARG * NIL NIL NIL VAR 28)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)) (28 (THING -) (CFORM FIN) :OBLIGATORY))
)
----------------------------------------------------------------------
[S-2]:
----------------------------------------------------------------------
;; Grid: 29.5.b#1#_exp_prop(que)#

(DEFINE-WORD
  :DEF_WORD "saber"
  :GLOSS "know"
  :CLASS "29.5.b"
  :WN_SENSE (("1.5" 333362 333754) ("1.6" 401762 402210))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT CIRCUMSTANTIAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2) (:ARG * NIL NIL NIL VAR 27)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)) (27 (THING -) (CFORM FIN)))
  :COLLOCATIONS ((27 "que"))
)
----------------------------------------------------------------------

[S-3]:
----------------------------------------------------------------------
;; Grid: 29.5.c.i#1#_exp_perc(de)#

(DEFINE-WORD
  :DEF_WORD "saber"
  :GLOSS "know"
  :CLASS "29.5.c.i"
  :WN_SENSE (("1.5" 333362) ("1.6" 401762))
```

```
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG * POSITION [ABOUT] PERCEPTUAL NIL 7
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG NIL THING NIL NIL VAR 8)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)))
)
----------------------------------------------------------------------

[S-4]:
----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "conocer"
  :GLOSS "know"
  :CLASS "29.5.c.ii"
  :WN_SENSE (("1.5" 333362) ("1.6" 401762))
  :LANGUAGE SPANISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION [ABOUT] PERCEPTUAL NIL 7
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * THING NIL NIL VAR 8)))
         (:MOD NIL MANNER KNOW+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)))
```

5.1.1.2 Sense distinctions according to semantic types of context

a) **[Encender]**, in Spanish, can be translated into English as "to light", "to switch on" or "to set on fire", depending on the semantic type of the object.

E: to light a candle/cigarette

S: encender una vela / cigarrillo

E: to switch on the radio, tv

S: encender la radio, la tele

E: to set on fire/ignite stubble

S: encender el rastrojo

b) The adjective **[groβ]** in German, can either be translated as "large" or "big" in English, or as "grande" or "grosse" in French, depending on context.

G: ein groβes Zimmer
F: une grande chambre
E: a large room

G: ein groβes Auto
F: une grosse voiture
E: a big car

c) Translations of the English verb **[shake]** depend on its argument type (e.g., abstract such as 'ideas' versus concrete such as 'bag') or whether the shake refers to an internal motion of the subject ('tremble' sense). The same for **[break].**

E: shake a bag

P: sacudir um saco

K: na-nun kapang-ul huntul-ess-ta

(I-Top bag-Acc shake-Past-Decl)

I: agitare / scuotere una borsa (*'Maria agitava* / *scuoteva una borsa'*)

S: agitar una bolsa

E: his ideas shook me

P: suas ideias me abalaram

I: turbare / colpire

(*'Le sue idee mi hanno turbato* / *colpito'*)

S: conmocionar (*sus ideas me conmocionaron*)

E: My hands shook

K: na-nun tali-ka tteli-ess-ta

(I-Top legs-Nom shake-Past-Decl)

I: tremare (*'Le mie mani tremano'*)

S: temblar (*mis manos tiemblan*)

P: tremer

[break]

E: John broke the window

K: Chelswu-ka changmwun-ul kkayttuly-ess-ta

(Chelswu-Nom window-Acc break-Past-Decl)

I: Gianni ha rotto / infranto la finestra

S: romper (Juan rompió la ventana)

E: John broke the law

K: Chelswu-ka pep-ul eky-ess-ta

(Chelswu-Nom law-Acc break-Past-Decl)

I: Gianni ha violato la legge

S: violar / quebrantar (Juan violó /quebrantó la ley)

E: The car broke

I: La macchina si e' rotta /guastata

The same action from a different perspective, sometime requiring changes to arguments:

[bring/take]

E & S: bring/traer (agent,patient,destination=here)

take/llevar (agent,patient,destination=there)

C: portar (agent,patient,destination=here&there)

E: bring/carry the books home

S: traer los libros a casa

C: portar els llibres a casa

E: take/carry the books to the school

S: llevar los libros a la escuela

C: portar els llibres a l'escola

5.1.1.2.1 Sense distinctions according to semantic types in Collins Gem

word :big

translation 1 : grand(e)

translation 2 : gros(se)

word : shake

translation 1 : secouer

translation 2 : agiter

translation 3 : ébranler

translation 4 : trembler

(Morpho)Syntactic constraint on translation 1 : subcategorization frame (vt)

(Morpho)Syntactic constraint on translation 3 : subcategorization frame (vt)

(Morpho)Syntactic constraint on translation 4 : subcategorization frame (vi)

Semantic constraint on translation 3 : domain (house, confidence)

5.1.1.2.2 Sense distinctions according to semantic types in PAROLE-Simple

Relevant Information in P-S: (i.) Template_type (link to a node in the ontology); (ii.) Syntactic Unit; (iii.) Predicative_Representation

Italian	(a.) *portare* (to carry / bring something): "Gianni portò la cravatta a Maria" (John brought the book to Mary) (b.) *portare* (to wear something) "Gianni porta la cravatta" (John wears the tie)
Analysis in P-S	**SemU**: *portare* (a.) → **SemU**: *bring* **Template_type**: Cause_change_of_location → **Template_type**: Cause_change_of_location **Synt. Construction**: → **Synt. Construction**: pos1 = NP; pos2 = NP; pos3 = a_PP → pos1 = NP; pos2 = NP; pos3 = a_PP **Arg. Struct.**: (<arg0>,<arg1><arg2>) → **Arg. Struct.**: (<arg0>,<arg1><arg2>) **SemU**: *portare* (b.) → **SemU**: *wear* **Template_type**: Relational_act → **Template_type**: Relational_act **Synt Construction**: → **Synt Construction**: pos1 = NP; pos2 = NP → pos1 = NP; pos2 = NP **Arg. Struct.**: (<arg0>,<arg1: Clothes>) → **Arg. Struct.**: (<arg0>,<arg1: Clothes>)

Relevant Information in P-S: (i.) Template_type (link to a node in the ontology); (ii.) Domain; (iii.) Qualia Structure

Italian	(a.) *colpire* (to hit somebody with something): "Gianni mi ha colpito con il Martello" (John hit me with the hammer) (b.) *colpire* (to impress somebody) "Il film ha colpito Maria" (The movie impressed Mary) (c.) *colpire* (to damage something) "Il terremoto ha colpito la Cina" (The quake damaged the China)

Analysis in P-S	**SemU**: *colpire* (a.) →	**SemU**: *hit* **Template_type**: Relational_act **Constitutive**: Contact=yes
	SemU: *colpire* (b.) →	**SemU**: *impress* **Template_type**: Cause_Experience **Domain**: Psychology
	SemU: *colpire* (c.) →	**SemU**: *damage* **Template_type**: Relational_Act

5.1.1.2.3 Sense distinctions according to semantic types in Euro(/Ital)WordNet

[Shake]

E: shake a bag

I: agitare, scuotere una borsa

{**Agitare**, riscuotere, menare, scuotere, vibrare, dimenare}

Definition: *muovere in qua e il là*

Has_Hyperonym: muovere

Top Concept: Cause, Location, Physical

EQ_SYNONYMY relation with:

{**shake**, agitate}

Definition: *move back and forth*;

Has_Hyperonym: move

Top Concept: Cause, Location, Physical

E: my hands shook

I: le mie mani tremano

{**tremare**, vibrare}

Top Concept: Location, Dynamic

EQ_NEAR_SYNONYMY relation with:

{oscillate, vibrate}

Definition: *move or swing from side to side regularly*

And EQ_NEAR_SYNONYMY relation with:

{tremble, **shake**, didder}

Definition*: move with a tremor*

E: his ideas shook me

I: le sue idee mi hanno colpito

{**colpire**, scioccare, impressionare}

Top Concept: Cause

EQ_NEAR_SYNONYMY relation with:

{**shock**, stun, floor, ball over, tack aback, blow out of the water}

Definition: *surprise greatly*;

Top Concept: Cause

5.1.1.2.4 Sense distinctions according to semantic types in EUROTRA

```
lex47 = {e_lu=disponer,e_isrno='3'} => {gb_lu=prepare, gb_rno=1}.
lex48 = {e_lu=disponer,e_isrno='2'} => {gb_lu=arrange, gb_rno=1}.
lex49 = {e_lu=disponer,e_isrno='1'} => {gb_lu=have, gb_rno=1}.
```

These elements are references to the full monolingual entries which contain more information. This information is used for disambiguation and takes into account, for instance, the semantic typing of the arguments (disponer_3 and disponer_2 have an arg1 which must be human, while disponer_1 has to be a concrete).

```
disponer_3 =
{cat=v,e_lu=disponer,e_isrno='3',e_isframe=arg1_2,
e_pformarg1=nil,e_pformarg2=nil,e_pformarg3=nil,e_pformarg4=nil,
p1type=nil,p2type=nil,
semarg1=hum,semarg2=conc,semarg3=nil,semarg4=nil,
e_vtype=main,vfeat=nstat,attype=nil,instrumental=yes,erg=no,
term='0', source=ttt,
definition='Colocar, poner las cosas en orden y situación
conveniente.',
example='El mayordomo ha dispuesto las habitaciones para los
invitados.'
%% xread_no='3'
%% lex_name=disponer}.

disponer_2 =
{cat=v,e_lu=disponer,e_isrno='2',e_isframe=arg1_2,
e_pformarg1=nil,e_pformarg2=nil,e_pformarg3=nil,e_pformarg4=nil,
p1type=nil,p2type=nil,
semarg1=hum,semarg2=sit,semarg3=nil,semarg4=nil,
e_vtype=main,vfeat=nstat,attype=vol,instrumental=no,erg=no,
term='0', source=ttt,
definition='Deliberar, determinar, mandar lo que ha de hacerse',
example='El gobierno ha dispuesto el envío de barcos al Golfo . La
Unesco ha dispuesto enviar ayuda humanitaria a la India'
%% xread_no='2'
%% lex_name=disponer}.
```

In the case of *disponer*_1 the main distinctive feature is the presence of a bound PP as arg2 marked with the preposition 'de' (of).

```
disponer_1 =
{cat=v,e_lu=disponer,e_isrno='1',e_isframe=arg1_2,
e_pformarg1=nil,e_pformarg2=de,e_pformarg3=nil,e_pformarg4=nil,
p1type=nil,p2type=nil,
semarg1=conc,semarg2=ent,semarg3=nil,semarg4=nil,
e_vtype=main,vfeat=stat,attype=nil,instrumental=no,erg=no,
term='0', source=ttt,
definition='Valerse de una persona o cosa, tenerla o utilizarla
por
```

```
suya. En sentido más amplio, tener.',
example='J y M disponen de poco tiempo para preparar el viaje. La
casa dispone de tres habitaciones para invitados.'
%% xread_no='1'
%% lex_name=disponer}.
```

In the case of nouns, such as *management*, direct reference to the semantic typing was also used for lexical selection. It has to be taken into account that no full agreement about the set of semantic typing features was achieved in Eurotra.

```
tlex7 = {gb_lu=management,rsf_human=yes} =>
{e_lu=dirección,sem=org}.
tlex8 = {gb_lu=management,rsf_human=no} => {e_lu=gestión}.


management={gb_lu=management,cat=n,gb_rno=1,morph_source=verbal,nc
lass=common,n_morphol=none,rsf_human=yes,rsf_loc=space,rsf_coll=ye
s,det_use=always_the,v_agr=sing,plurality=no_pl,ers_frame=none,t=n
o,wh=no,source=tc,person=third}.

management={gb_lu=management,cat=n,gb_rno=2,morph_source=verbal,nc
lass=common,n_morphol=none,rsf_human=no,rsf_loc=none,rsf_coll=no,d
et_use=never_det,v_agr=sing,plurality=no_pl,ers_frame=subj_objnp,t
=no,wh=no,source=tc,person=third}.
```

5.1.1.2.5 Sense distinctions according to semantic types in SYSTRAN

gross .if_modifies_noun+LOCATION then trnsl. EN “big”

shake .if_ semantic_object_is_noun+CONCRETE+NOT_ANIMATE then trsnsl ….

This takes care of : They shook the bag

The bag was shaken

.. The bag, shaken by the man, broke

… The shaken bag

Shake .if_no_object and .if_subject_is_noun+HUMAN then trnsl …

The man shook.

Shake .if_no_object and .if_subject_is_noun+DEVICE then trnsl …

His hands shook.

The engine shook.

For bring and take, SYSTRAN would have two similar expressions and give the same CAT trnsl for both

5.1.1.2.6 Sense distinctions according to semantic types in Lexical Conceptual Structure Lexicon

Asymmetrical hyponyms

[bring/take]

E: bring/carry the books home [E-5][E-6]
S: traer los libros a casa
CAT: portar els llibres a casa

E: take/carry the books to the school [E-6][E-7]
S: llevar los libros a la escuela
CAT: portar els llibres a l'escola

```
[E-5]
----------------------------------------------------------------------
;; Grid: 11.3#1#_ag_th,src(from),goal(to)#

(DEFINE-WORD
  :DEF_WORD "bring"
  :CLASS "11.3"
  :WN_SENSE (("1.5" 1188762 824200 823804 1271735)
             ("1.6" 1422262 1422262 982468 1527059))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL EVENT CAUSE NIL NIL 37
        ((:SUB * THING NIL NIL VAR 1)
         (:ARG NIL EVENT GO LOCATIONAL NIL 38
          ((:SUB * THING NIL NIL VAR 2)
           (:ARG * PATH TO LOCATIONAL NIL 5
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION AT LOCATIONAL NIL 39
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 6)))))
           (:ARG * PATH FROM LOCATIONAL NIL 3
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION AT LOCATIONAL NIL 40
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 4)))))))))
```

```
        (:MOD NIL MANNER BRING+INGLY NIL NIL 26)))
  :VAR_SPEC ((5 :OPTIONAL) (3 :OPTIONAL) (1 (ANIMATE +)))
)
---------------------------------------------------------------------

[E-6]
---------------------------------------------------------------------
;; Grid: 11.4.ii#1#_ag_th,src(from),goal(to)#

(DEFINE-WORD
  :DEF_WORD "carry"
  :CLASS "11.4.ii"
  :WN_SENSE (("1.5" 834152 1537537 1190169)
             ("1.6" 994853 1855700 1424107))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL EVENT CAUSE NIL NIL 37
        ((:SUB * THING NIL NIL VAR 1)
         (:ARG NIL EVENT GO LOCATIONAL NIL 38
          ((:SUB * THING NIL NIL VAR 2)
           (:ARG * PATH [TO] LOCATIONAL NIL 5
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION AT LOCATIONAL NIL 39
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 6)))))
           (:ARG * PATH [FROM] LOCATIONAL NIL 3
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION AT LOCATIONAL NIL 40
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 4)))))))
         (:MOD NIL MANNER CARRY+INGLY NIL NIL 26)))
  :VAR_SPEC ((5 :OPTIONAL) (3 :OPTIONAL) (1 (ANIMATE +)))
)
---------------------------------------------------------------------

[E-7]
---------------------------------------------------------------------
;; Grid: 11.3#1#_ag_th,src(from),goal(to)#

(DEFINE-WORD
  :DEF_WORD "take"
  :CLASS "11.3"
  :WN_SENSE (("1.5" 691086 379073 1258879 104355 1259481 1537537
              1257967 824200 620792)
             ("1.6" 826635 455018 1510674 118898 1511279 1855700
              1509715 1422262 744637))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL EVENT CAUSE NIL NIL 37
        ((:SUB * THING NIL NIL VAR 1)
         (:ARG NIL EVENT GO LOCATIONAL NIL 38
          ((:SUB * THING NIL NIL VAR 2)
           (:ARG * PATH TO LOCATIONAL NIL 5
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION AT LOCATIONAL NIL 39
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 6)))))
           (:ARG * PATH FROM LOCATIONAL NIL 3
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION AT LOCATIONAL NIL 40
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 4)))))))
         (:MOD NIL MANNER TAKE+INGLY NIL NIL 26)))
  :VAR_SPEC ((5 :OPTIONAL) (3 :OPTIONAL) (1 (ANIMATE +)))
```

5.1.1.3 Senses according to domain terms

a) **[File]** in English is polysemous, but in German it is translated as either "Feile" in the general domain or as "Datei" when in the computer domain.

E: file

G: Feile (general domain)

E: file

G: Datei (computer domain)

b) The two different senses of **[mouse]** (a homonym in English) are translated as two different lexical items in Italian, according to domain.

[mouse] Homonyms in English

E: Mouse

I: mouse (computer)

E: Mouse

I: topo (zool.)

5.1.1.3.1 Senses according to Domain terms in Collins Gem

word : avocat

translation 1 : barrister

translation 2 : avocado

Semantic constraint on translation 1 : domain

Semantic constraint on translation 2 : domain

5.1.1.3.2 Senses according to Domain terms in PAROLE-Simple

Sense distinctions reflected in terms of domain and semantic type

Relevant Information in P-S: (i.) Template_type (link to a node in the ontology); (ii.) Domain

Example	(a.) *mouse* (a type of animal) (It. *topo*) (b.) *mouse* (pointing device for computers) (It. *mouse*)
Analysis in P-S	**SemU**: *mouse* (a.) → **SemU**: *topo* **Template_type**: Animal → **Template_type**: Animal **Domain**: Zoology → **Domain**: Zoology **SemU**: *mouse* (b.) → **SemU**: *mouse* **Template_type**: Instrument → **Template_type**: Instrument **Domain**: Computing → **Domain**: Computing

5.1.1.3.3 Senses according to Domain terms in Euro(/Ital)WordNet

[**file**] polysemous in English

E: file

I: schedario (general domain)

{schedario, clasellario, classificatore}

EQ_SYNONYMY relation with:

{**file**, file cabinet, filing cabinet}

Definition: *a container for keeping papers in order*

E: file

I: file (computer domain)

{**file**, documento}

DOMAIN: Computer

EQ_SYNONYMY relation with:

{file, data file}

[**mouse**]homonyms in English

E: Mouse

I: Mouse

{**mouse**}

DOMAIN: Computer

EQ_SYNONYMY relation with:

{mouse}

Definition: *a hand operated device that moves the cursor on a computer screen*

E: Mouse

I: Topo

{topo, sorcio}

EQ_SYNONYMY relation with:

{mouse}

Definition: *small rodents*

5.1.1.3.4 Senses according to Domain terms in SYSTRAN

SYSTRAN distinguishes domain-specific translations in the Stem Dictionary.

e.g. the entry for file has a transl for “TG= technical” DE “Feile”

and TG=computer” as DE “Datei”

similarly for Mouse

Of course this is not sufficient to keep the two meanings apart.

Therefore, there will also be entries in the Expression Dictionary

e.g. mouse pad

click … mouse

etc.

5.1.1.4 Number (nb)

There are some cases where the languages we treat differ with respect to number.

5.1.1.4.1 Differences respect to number in EUROTRA

DA: USA (sing)	ES: EEUU (plu)
EN: persons	ES: gente (sing)

In these cases the value of number has to be changed, both on the leaf node and on the np-node. The leaf rule is as follows:

```
b10b = {cat=n,da_lu=usa,nb=sing} => {e_lu=eeuu,nb=plu}.
```

5.1.2 Examples of differences in predicate argument structure

It is often the case that a translation has inverted arguments mapping or differences in the syntactic structure as shown in the examples below:

E: I like Mary

F: Marie plaît `a moi. (Marie me plait)

(Mary is pleasing to me)

G: Maria gefällt mir

I: A me piace Maria

E: I miss Mary

F: Marie me manque

G: Maria fehlt mir

I: A me manca Maria

5.1.2.1.1 Inverted argument mappings in Collins Gem

word : manquer

multiword : il/cela nous manque

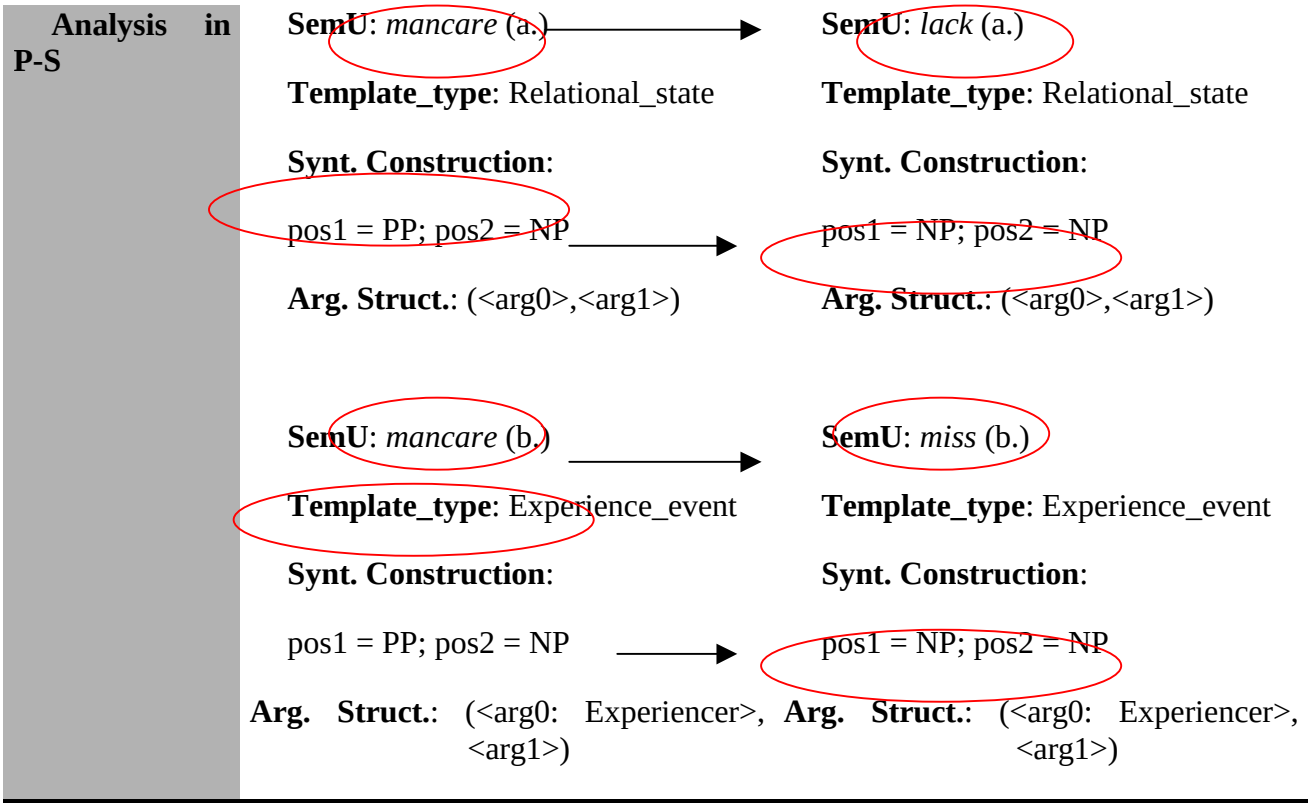
translation : I miss him/this

5.1.2.1.2 Inverted argument mappings in PAROLE-Simple

Sense distinctions reflected in terms of semantic type, syntactic frames and argument structure:

Relevant Information in P-S: (i.) Template_type (link to a node in the ontology); (ii.) Syntactic Unit; (iii.) Predicative_Representation

Italian	(a.) *mancare* (to lack something): "A me mancano soldi" (I lack money) (b.) *mancare* (to miss somebody) "A me manca Maria" (I miss Mary)



5.1.2.1.3 Inverted arguments mapping in EUROTRA

Role changes cannot be performed elegantly. The role a phrase plays depends on the subcategorisation frame of its governor. Such role changes are never general but lexically dependant. The necessary information to contextualise a rule - that is the lu of the governor- is present at a higher node. The only way to change the role was by means of a structural rule , deleting the phrase, whose role changes from one language to the other, at the left-hand side of the rule and recreating it on the right-hand side. The information of the phrase which is copied to the TL has to be explicitly saved with variables.

examples:

EN: I like Mary
arg1 arg2

ES: Me gusta María
arg2 arg1

rule:
:b:

```
tlike =  S:{cat=s}[V:{cat=v,gb_lu=like},
                   ~:{cat=np,role=arg1}
                      [N:{}],
                   ~:{cat=np,role=arg2}
                      [N2:{}]]
```

```
=>
          S:{cat=s}<V:{cat=v,e_lu=gustar},
                     {cat=np,role=arg1}
                            <N2>,
                     {cat=np,role=arg2}
                            <N>>.
```

5.1.2.1.4 Inverted arguments mappings in SYSTRAN

SYSTRAN handles the following type (subject – dativeObject swapping) by attaching a special code to the translation

E: “like”

Trsl G: “gefallen + DATSUB”

Trsl FR “plaire + DATSUB”

This code triggers a program which performs all the necessary transformations.

5.1.2.1.5 Inverted arguments mappings in Lexical Conceptual Structure Lexicon

E: I like Mary [E-8]
S: Maria me gusta [S-5]

```
[E-8]:
------------------------------------------------------------------------
;; Grid: 31.2.a#1#_exp_perc,purp(for),mod-pred(as)#

(DEFINE-WORD
  :DEF_WORD "like"
  :CLASS "31.2.a"
  :WN_SENSE (("1.5" 1012304 1012137) ("1.6" 1213391 1213205))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT PERCEPTUAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * THING NIL NIL VAR 8)))
         (:MOD * POSITION FOR INTENTIONAL NIL 21
```

```
          ((:SUB NIL STATE *HEAD* NIL NIL 39)
           (:ARG NIL THING NIL NIL VAR 22)))
         (:MOD * POSITION AS IDENTIFICATIONAL NIL 29
          ((:SUB NIL STATE *HEAD* NIL NIL 40)
           (:ARG NIL THING NIL NIL VAR 30)))
         (:MOD NIL MANNER LIKE+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)) (8 (ANIMATE +)) (21 :OPTIONAL) (29 :OPTIONAL))
)
-----------------------------------------------------------------------

[S-5]:
-----------------------------------------------------------------------
;; Grid: 31.2.a#1#_exp_perc,purp(por,para),mod-pred(como)#

(DEFINE-WORD
  :DEF_WORD "gustar"
  :GLOSS "like"
  :CLASS "31.2.a"
  :WN_SENSE (("1.5" 1012304 1012137) ("1.6" 1213391 1213205))
  :LANGUAGE SPANISH
  :LCS (:ROOT NIL STATE BE PERCEPTUAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT PERCEPTUAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * THING NIL NIL VAR 8)))
         (:MOD * POSITION FOR INTENTIONAL NIL 21
          ((:SUB NIL STATE *HEAD* NIL NIL 39)
           (:ARG NIL THING NIL NIL VAR 22)))
         (:MOD * POSITION AS IDENTIFICATIONAL NIL 29
          ((:SUB NIL STATE *HEAD* NIL NIL 40)
           (:ARG NIL THING NIL NIL VAR 30)))
         (:MOD NIL MANNER LIKE+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +) :INT) (8 (ANIMATE +) :EXT)
            (21 :OPTIONAL) (29 :OPTIONAL))
)
```

5.1.3 Examples involving more than a single lexical item

5.1.3.1 Predicative nominals that are predicative adjectives in another language, and/or that take different auxiliaries (Categorial)

E: I am hungry

F: J'ai faim

G: Ich habe Hunger

(I have hunger)

I: a. Sono affamato (adj.)

b. ho fame (N.)

S: a. tener hambre (N.)

b. estar hambriento /sediento (adj)

E: That problem is important

K: ku mwuncey-ka cwungyoha-ta.

(that problem-Nom important-Decl)

5.1.3.1.1 Categorials in Collins Gem

word : hungry

multiword : to be hungry

translation : avoir faim

5.1.3.1.2 Categorials in EUROTRA

Category changes can occur at different levels. We will only deal with those category changes which are performed from leaf node to leaf node, i.e. where the structure as such can be mantained but the feature containing information about the syntactic category has to be changed from SL to TL.

examples:

DA:nogen (cat=adj) ES:algún (cat=quant)

rules:

```
tnogen1 = {cat=adj,dalu=nogen}=>{cat=quant,e_lu=algún}.
```

5.1.3.1.3 Categorials in SYSTRAN

SYSTRAN simply translates the words by the different category

E.g be .if_pred.adjective_is “hungry”

Trsl be as FR “avoir”

Trsl hungry as FR “faim”

There are only a limited number of expressions of this type in the western European languages. It would certainly also be possible to change the categories to the ones required by the target language.

For the Korean example, a part of the transfer program is executed that transfers all verb information from the copula to the adjective and makes the (conjugable) Korean adjective into the predicate of the translated sentence. Nothing special is done in the dictionary.

5.1.3.1.4 Categorials in Lexical Conceptual Structure Lexicon

E: I am hungry [E-9] from (Dorr,1993)
G: Ich habe Hunger [G-1] from (Dorr,1993)

```
[E-9]
----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "be"
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE IDENTIFICATIONAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT IDENTIFICATIONAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * PROPERTY NIL NIL VAR 8)))))
  :VAR_SPEC ((2 (HUMAN +)) (8 (ANIMATE +)))
```

```
)

(DEFINE-WORD
  :DEF_WORD "hungry"
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL PROPERTY HUNGRY+/P NIL NIL 0)
)
-----------------------------------------------------------------------

[G-1]
-----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "haben"
  :LANGUAGE GERMAN
  :LCS (:ROOT NIL STATE BE IDENTIFICATIONAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT IDENTIFICATIONAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * PROPERTY NIL NIL VAR 8)))))
  :VAR_SPEC ((2 (HUMAN +)) (8 (:CAT N)))
)

(DEFINE-WORD
  :DEF_WORD "Hunger"
  :LANGUAGE GERMAN
  :LCS (:ROOT NIL PROPERTY HUNGRY+/P NIL NIL 0)
)


NOTE: possessional haben would look like this:

(DEFINE-WORD
  :DEF_WORD "haben"
  :LANGUAGE GERMAN
  :LCS (:ROOT NIL STATE BE POSSESSIONAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT POSSESSIONAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG * THING NIL NIL VAR 8)))))
  :VAR_SPEC ((2 (HUMAN +)))
)
```

5.1.3.2 Conflational: a single word in one language is a phrase in another

E: farmer's wife

F: fermiere

I: fattora (rare)

F: Il a pris sa retraite

E: He retired

G: Er ist in den Ruhestand getreten

5.1.3.2.1 Conflationals in Collins Gem

word : fermier, ière

translation 1 : farmer

translation 2 : farmer's wife

Syntactic constraint on translation 1 : morphosyntactic (nm)

Syntactic constraint on translation 2 : morphosyntactic (nf)

word : retraite

multiword : prendre sa retraite

translation 1 : to retire

5.1.3.2.2 Conflationals in Euro(/Ital)WordNet

E: face powder

I: cipria

{**cipria**}

Definition: *sottile polvere per truccare il viso*

Has_Hyperonym: cosmetico

EQ_SYNONYMY relation with:

{face powder}

Has_Hyperonym: cosmetics

5.1.3.2.3 Conflationals in SYSTRAN

SYSTRAN allows translation of one word by several and vice versa.

In the first case, special dictionary codes are used to indicate which and how many of the words in the translation need to be inflected. There are also codes to indicate the order of words in the translation.

E: farmer's wife

Trsl F: "fermiere"

E: retire

Trsl DE " in den Ruhestand treten" (+ a code that indicates other word orders; e.g. er tritt in den Ruhestand)

5.1.3.3 Argument incorporation differences: some arguments in one language are incorporated into the head in the other language

E: to funnel

P: colocar com um funil

(put with a funnel)

I: versare con l'imbuto

5.1.3.3.1 Argument incorporation differences in Collins Gem

type of phenomena : multiword constructions

subtype : argument incorporation

word : funnel, bicycle, etc.

5.1.3.3.2 Argument incorporation differences in Euro(/Ital)WordNet

E: to funnel

I: versare con l'imbuto

{**imbuto**}

Definition: *strumento di forma conica per versare liquidi all'interno di recipienti*

Has_Hyperonym: strumento

EQ_SYNONYMY relation with:

{**Funnel**}

Definition: *a conically shaped utensil*

EQ_ROLE relation with:

{to funnel}

Definition: *pour through a funnel*

5.1.3.3.3 Argument incorporation differences in SYSTRAN

For SYSTRAN:similar to the above examples, except here it is often necessary to indicate which word is the head word and whether the direct object must be inserted in the coded expression. The code INSOBJ is used for this.

e.g. EN funnel (verb)

Trsl colocare com um funil +WN1=verb+ INSOBJ

5.1.3.3.4 Argument incorporation differences in Lexical Conceptual Structure Lexicon

E: to funnel
S: encauzar con un embudo

```
[E-10]
----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "funnel"
  :GLOSS "funnel"
  :CLASS "9.3.a"
  :WN_SENSE (830384)
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL EVENT CAUSE NIL NIL 36
        ((:SUB * THING NIL NIL VAR 1)
         (:ARG NIL EVENT GO LOCATIONAL NIL 37
          ((:SUB * THING NIL NIL VAR 2)
           (:ARG * PATH [TOWARD] LOCATIONAL NIL 5
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION [IN] LOCATIONAL NIL 38
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 6)))))))
         (:MOD * POSITION WITH INSTRUMENTAL NIL 19
          ((:SUB NIL EVENT *HEAD* NIL NIL 39)
           (:ARG NIL THING FUNNEL+ER NIL NIL 20)))))
  :VAR_SPEC ((1 (ANIMATE +)))
)
----------------------------------------------------------------------

[S-6]
----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "encauzar"
  :GLOSS "funnel"
  :CLASS "9.3.a"
  :WN_SENSE (("1.5" 830384) ("1.6" 990205))
  :LANGUAGE SPANISH
  :LCS (:ROOT NIL EVENT CAUSE NIL NIL 37
        ((:SUB * THING NIL NIL VAR 1)
         (:ARG NIL EVENT GO LOCATIONAL NIL 38
          ((:SUB * THING NIL NIL VAR 2)
           (:ARG * PATH [TOWARD] LOCATIONAL NIL 5
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL POSITION [IN] LOCATIONAL NIL 39
              ((:SUB NIL THING NIL NIL VAR 2)
               (:ARG NIL THING NIL NIL VAR 6)))))))
         (:MOD * POSITION WITH INSTRUMENTAL NIL 19
          ((:SUB NIL EVENT *HEAD* NIL NIL 40)
           (:ARG NIL THING FUNNEL+ER NIL NIL 20)))))
```

```
:VAR_SPEC ((5 :OPTIONAL) (1 (ANIMATE +)))
```

5.1.3.4 Head switching: some examples of demotional and promotional phenomena, when modifiers in one language may become matrix verbs in another and vice-versa.

E: I like to eat

G: Ich esse gern

(I eat likingly)

I: a. Mi piace mangiare (this has only an habitual, generic meaning: i.e. eating is a favourite passion of the speaker)

b. Mangio volentieri (Besides a generic reading, also "I feel like eating")

E: She smiled her thanks

F: Elle remercia d'un sourire

G: Sie bedankte sich mit einem Lächeln

5.1.3.4.1 Head switching in Collins Gem

word : smile

multiword : to smile her thanks

5.1.3.4.2 Head Switching in SYSTRAN

SYSTRAN: These are coded as specific expressions (a more generalized solution could be implemented, but hasn't been)

EN like .if_governs_inf="eat"

Trsl like as DE "gern essen"

EN like .if_governs_inf=”drink”

Trsl like as DE “gern trinken”

(other word order “isst gern” is generated automatically for German; no special dictionary code is needed)

5.1.3.4.3 Head Switching in Lexical Conceptual Structure

E: I like to eat [E-11] from (Dorr,1993)
G: Ich esse gern [G-2] from (Dorr,1993)

```
[E-11]
----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "like"
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL STATE BE CIRCUMSTANTIAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT CIRCUMSTANTIAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2) (:ARG * NIL NIL NIL VAR 27)))
         (:MOD NIL MANNER LIKE+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)) (27 (THING -) (CFORM INF)))
  :COLLOCATIONS ((27 "to"))
)
----------------------------------------------------------------------

[G-2]
----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "gern"
  :GLOSS "like"
  :LANGUAGE GERMAN
  :LCS (:ROOT NIL STATE BE CIRCUMSTANTIAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL POSITION AT CIRCUMSTANTIAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2) (:ARG * NIL NIL NIL VAR 27)))
         (:MOD NIL MANNER LIKE+INGLY NIL NIL 26)))
  :VAR_SPEC ((2 (HUMAN +)) (27 (THING -) :DEMOTE))
)
```

5.1.3.4.4 Path verbs

E: John swam across the river

K: Chelswu-ka swuyenghase kang-ul kenne-ss-ta

(Chelswu-Nom swim river-Acc cross-Past-Decl)

F: traverser `a la nage

(cross by swimming)

I: attraversare a nuoto

S: atravesar a nado

E: I emailed the note to John

I: Ho spedito il messaggio a Gianni per e-mail

S: mandar un mensaje por correo electrónico

5.1.3.4.4.1 Path Verbs in Collins Gem

word :swim

translation 1: nager

translation 2 : traverser (à la nage)

syntactic constraint on translation 1 : syntactic (vi)

syntactic constraint on translation 2 : syntactic (vt) type of phenomena : multiword constructions

5.1.3.4.4.2 Path Verbs in SYSTRAN

SYSTRAN expression for these would be :

EN swim .if_prep_compl=“across” and .if_prep_object =river,lake,or any WATERBED”

Then trsl swim as FR “traverser a la nage” (verb=WN1 + INSOBJ)

5.1.3.4.4.3 Path Verbs in Lexical Conceptual Structure Lexicon

E: swim across [E-12]
S: atravesar a nado [S-7]

```
[E-12]
-----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "swim"
  :CLASS "51.3.2.a.ii"
  :WN_SENSE (("1.5" 1116739 1084706) ("1.6" 1335172 1299337))
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL EVENT GO LOCATIONAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG * PATH FROM LOCATIONAL NIL 3
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG NIL POSITION [AT] LOCATIONAL NIL 38
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL THING NIL NIL VAR 4)))))
         (:ARG * PATH TO LOCATIONAL NIL 5
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG NIL POSITION [AT] LOCATIONAL NIL 39
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG NIL THING NIL NIL VAR 6)))))
         (:MOD NIL MANNER SWIM+INGLY NIL NIL 26)))
  :VAR_SPEC ((5 :OPTIONAL) (3 :OPTIONAL))
)

(DEFINE-WORD
  :DEF_WORD "across"
  :LANGUAGE ENGLISH
  :LCS (:ROOT NIL POSITION ACROSS LOCATIONAL NIL 0
        ((:SUB NIL NIL NIL NIL VAR 2) (:ARG NIL NIL NIL NIL VAR 11)))
  :VAR_SPEC ((0 (:CAT ADV)))
)
-----------------------------------------------------------------------

[S-7]
-----------------------------------------------------------------------
(DEFINE-WORD
  :DEF_WORD "atravesar"
  :GLOSS "cross"
  :CLASS "51.1.h"
  :WN_SENSE (("1.5" 1089601) ("1.6" 1304824))
  :LANGUAGE SPANISH
  :LCS (:ROOT NIL EVENT GO LOCATIONAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG NIL PATH TOWARD LOCATIONAL NIL 38
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG NIL POSITION ACROSS LOCATIONAL NIL 39
            ((:SUB NIL THING NIL NIL VAR 2)
             (:ARG * THING NIL NIL VAR 6)))))
         (:MOD NIL MANNER CROSS+INGLY NIL NIL 26)))
  :VAR_SPEC ((6 :OPTIONAL) (2 (ANIMATE +)))
)
```

```
(DEFINE-WORD
  :DEF_WORD "nadar"
  :GLOSS "swim"
  :CLASS "47.5.1.b"
  :WN_SENSE (("1.5" 1116739 1084706) ("1.6" 1335172 1299337))
  :LANGUAGE SPANISH
  :LCS (:ROOT NIL EVENT ACT LOCATIONAL NIL 37
        ((:SUB * THING NIL NIL VAR 2)
         (:ARG * POSITION [AT] LOCATIONAL NIL 10
          ((:SUB NIL THING NIL NIL VAR 2)
           (:ARG NIL THING NIL NIL VAR 11)))
         (:MOD NIL MANNER SWIM+INGLY NIL NIL 26)))
  :VAR_SPEC NIL
)
```

5.1.3.5 No literal translation, requires an entry in a phrasal lexicon

E: John broke into the room

S: Juan forz'o la entrada al cuarto

(John forced entry to the room)

I: Gianni fece irruzione / entrò con la forza nella stanza

E: shake hands

P: apertar m~aos

(squeeze hands)

K: na-nun John-kwa akswu-lul ha-yess-ta

(I-Top John-with hand_shake-Acc do-Past-Decl)

I: stringersi /darsi la mano

('Gianni e Mario si sono stretti / dati la mano')

S: darse la mano (Juan y María se dieron la mano)

5.1.3.5.1 No literal translation in Collins Gem

word : break

multiword : to break into

translation1 : s'introduire dans

Syntactic constraint on translation 1 : syntactic (vi)

Semantic constraint on translation 1 : domain (house)

5.1.3.5.2 No literal translation in SYSTRAN

SYSTRAN treats these similarly to the others above

EN break .if_prepos_complement= "into" and .if_prep_object is "room" or any other enclosed space

Then trsl break in ES "forzar la entrada a"

5.1.3.5.3 No literal translation in Lexical Conceptual Structure Lexicon

E: John broke into the room
S: Juan forz'o la entrada al cuarto
(John forced entry to the room)

(Dorr, 1993) describes an LCS treatment of break-into. The following picture describes how the LCS is decomposed into three languages: English, Spanish and Arabic.

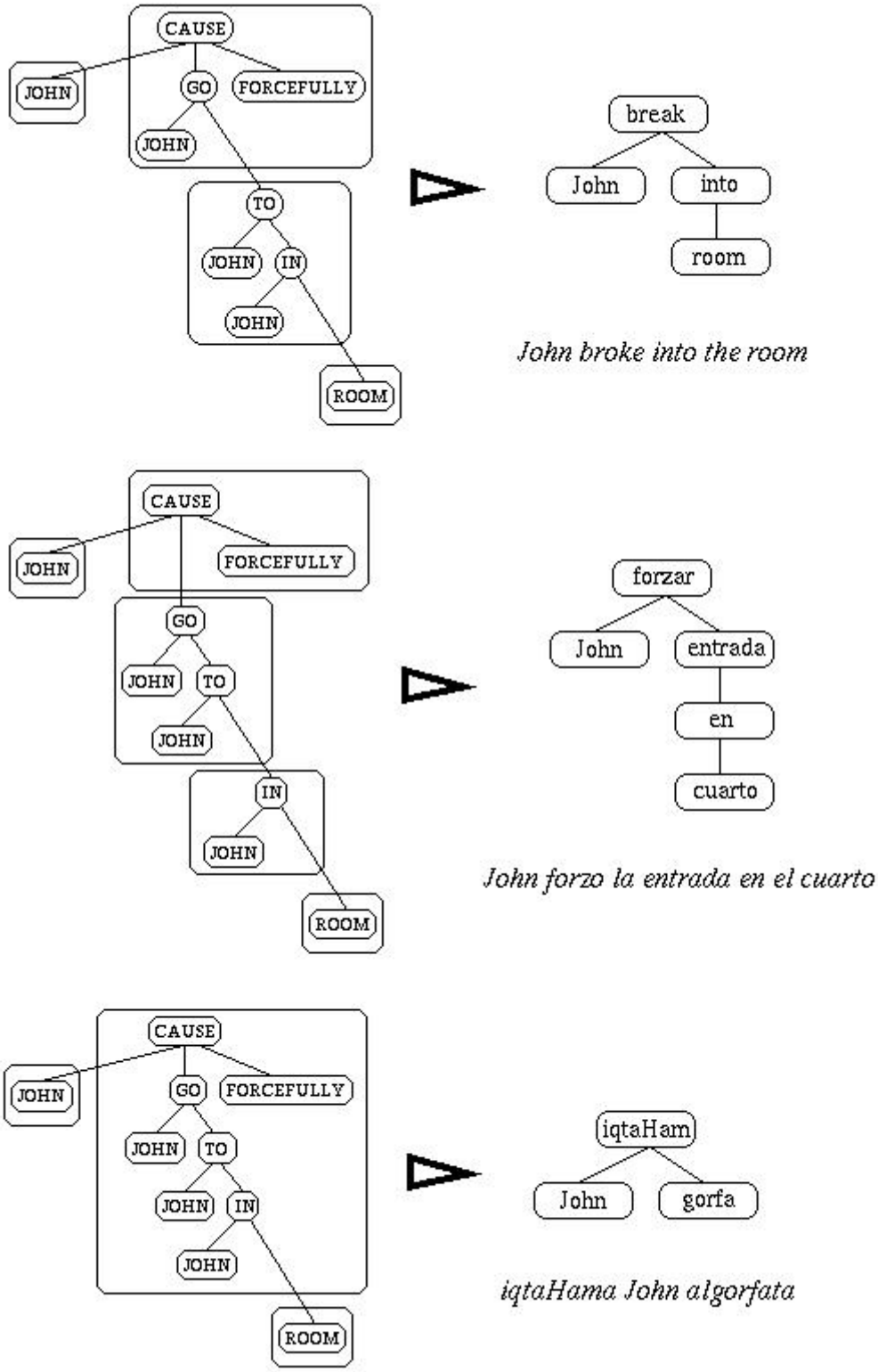


Fig. 20: LCS treatment of break-into

5.1.4 Multi-word constructions: idioms

5.1.4.1 Verb phrases

a) almost completely frozen (cannot be internally modified, don't allow passivization, singular/plural alternation, etc.)

E: let the cat out of the bag

E: John kicked the bucket

P: Jo~ao morreu

I: Gianni ha tirato le cuoia

b) somewhat modifiable

E: know/teach/learn the ropes

I: Vederne / dirne / passarne /farne di tutti i colori (to see / to tell / to go through /to do all sorts of things)

c) internal variable:

bound to subject:

E: blow one's stack

bound to non-subject NP

E: got someone's number (unbound)

d) light verb + NP - might be a single word in target language

compositional/semantically transparent

E: to take a trip

I: fare un viaggio

E: to have fun

S: divertirse

non-compositional:

E: take the words out of one's mouth

I: fare caso a (to notice)

either (polysemous):

E: The engine overstrained

K: encin-ey mwuli-ka ka-ss-ta

(engin-Nom overstrain-Nom go-Past-Decl)

I: prendere la mano (to get out of somebody's control)

5.1.4.1.1 Verb phrases in Collins Gem

word : fun

idiom : to have fun

translation : s’amuser

5.1.4.1.2 Verb phrases in Euro(/Ital)WordNet

E: indispose

I: fare star male

{fare star male}

Has_Hyperonym: rendere, fare

EQ_SYNONYMY relation with:

{indispose, cause to feel unwell}

Has_Hyperonym: Change

5.1.4.2 NP

a) non-compositional compounds:

E: stepping stone

E: straight arrow

I: testa di ponte (bridgehead)

I: muro di gomma (somebody that is totally indifferent)

5.1.4.2.1 NP in Collins

word : stone

idiom : stepping stone

5.1.4.2.2 NP in SYSTRAN

SYSTRAN dictionaries indicate the head word of both the source NP and its translation.

5.1.4.3 Clauses, sentences

E: when the cows come home

E: not a leg to stand on

F: Vas te faire cuire un oeuf

E: one's bark is worse than one's bite

5.1.4.3.1 Clauses in Collins Gem

word :oeuf

idiom :vas te faire cuire un oeuf

5.1.4.3.2 Clauses in SYSTRAN

Some of these types of expressions are coded as “non-variable idioms” in SYSTRAN dict., i.e. the entire phrase is replaced by the entire phrasal translation. This type is very rare in SYSTRAN dictionaries .

6 Towars Multilingual ISLE Lexical Entry

6.1 A first comparison of the surveyed resources

The ISLE survey highlights interesting aspects and points of view in the multifarious scenario of the bilingual resources that are currently available in the HLT community. In this final section, we attempt to illustrate some of these perspectives, trying to foreground the major tendencies and generalizations, so as to provide a first important bootstrap for the next phases of the ISLE work, i.e. the standardization proposal. The survey shows that existing lexical multilingual resources can be grouped in at least four classes:

1. machine-readable dictionaries (MRD);
2. general purpose computational lexicons (GPCL);
3. application-oriented computational lexicons (AOCL);
4. lexical data representation and interchange formats (LDRIF).

Although they differ under many respects, these resources also show a great amount of overlapping and reciprocal interactions, both on the content and on the representational levels, which deserve to be made explicit.

6.1.1 Machine-readable dictionaries

MRDs like the Collins, the Oxford-Hachette (§. 3.1.1) and the Van Dale (§. 3.1.2) represent the most classical resources for HLT systems. Being essentially developed for human users, they maintain most of the characteristic of traditional paper dictionaries, both in the general architecture, as well as in the way linguistic information is organized and encoded. In general, differently from computational lexicons, they lack an explicit representation of linguistic information such as inflectional class, obligatory complements, alternations, regular polysemy, etc. The characterization of lexical entries is mostly achieved through a rich array of examples. *Prima facie*, MRDs are fairly orthogonal with computational lexicons, they nevertheless represent important resources on their own for multilingual HLT. First of all, MRDs are widely used as input to build computational lexicons (both AOCL and GPCL), as shown in the cases of Microsoft (§. 3.3.7) and of the Collins Robert Semantic Lexical Database (§. 3.2.1). Dictionary definitions and translation examples are widely used to populate computational lexicons with crucial information, and they allow the lexical resource construction to be a truly dynamic process. Secondly, although human user oriented, the structure of multilingual MRDs provide useful insights and inputs to the process of computational lexicon design. While many computational lexical databases try to make explicit large amounts of usually implicit lexical knowledge, MRDs show the crucial importance of linguistic examples to establish translation equivalents, as well as provide the crucial support and background of the best lexicographic tradition.

6.1.2 General purpose computational lexicons

EuroWordNet/ItalWordNet (§. 3.2.3), PAROLE/SIMPLE (§. 3.2.4), and FrameNet (§.3.2.2) represent important instances of GPCLs. Differently from MRDs they aim at making explicit morphosyntactic, syntactic and semantic knowledge, partly through an extensive work of extraction from corpora. They have an inherent vocation towards application-independence, since they encode general linguistic knowledge, rather than being exclusively tailored to the specific needs of some particular applications. With this respect, they represent general models of lexical architecture strongly grounded on well-established theoretical frameworks, which provide the main representational backbone (e.g. the Generative Lexicon, Frame Semantics, etc.). As a consequence, while guaranteeing a high degree of reusability and generality, GPCLs need to be specifically customized to apply to particular domains. EuroWordNet/ItalWordNet provides an interesting example of a general lexicon, which also contains a domain specific instantiation.

Most of the existing GPCLs are essentially monolingual, although it has been shown that the linguistic information they encode can be extremely useful in multilingual environments, and actually multiligual links of simple types in some cases already exist (cf. EuroWordNet). The only exception is represented by The Collins-Robert Lexical Semantic DataBase which is truly bilingual and actually is also an important case of interaction with MRDs. Semantic information is extracted out of a MRD, and represented through Mel'chuk lexical functions.

Even within the general category of GPCLs, the surveyed resources show big differences. PAROLE/SIMPLE, for instance, provides a large bulk of information (practically the whole set of the EAGLES recommended information types) but lacks collocational information, as well as the representation of multiword expressions, although the SIMPLE model allows for their fast integration into the existing architecture. On the other hand, EuroWordNet/ItalWordNet is by its own vocation oriented towards a network representation of lexical semantic information, while lacking information for argument structure and syntax-semantic mapping. With this respect, EuroWordNet/ItalWordNet and PAROLE/SIMPLE represent an interesting example of complementary lexical architectures. Finally, FrameNet shows an important corpus-oriented vocation, paired with strong theoretical assumptions, and important synergies can be foreseen with a model like PAROLE/SIMPLE, together with prospective extensions to cover areas such as MWEs and multilinguality. The future integration of EuroWordNet/ItalWordNet, PAROLE/SIMPLE and FrameNet should thus be supported and fostered, so as to get at a more comprehensive model for GPCLs.

6.1.3 Application-oriented computational lexicons

As for AOCLs, the present survey has mostly focussed on resources for MT systems. The main reason is that MT provides very interesting examples of different styles of multilingual lexicons, due also to the crucial role of such resources in the high-demanding task of automatic translation. In this area, we find large lexicons which provide very complex methods and solutions to establish translation equivalents and complex lexical multilingual mappings. All the surveyed lexicons establish translations equivalents in terms of rich arrays of morphosyntactic information encoded in the lexical entries (e.g. subcategorization frames, etc.). Conversely, semantic information has so far a less central role, which is also reflected into its less wide encoding in the lexicons. While Microsoft and EDR (§. 3.3.3) have very complex and articulate semantic components, semantic information are preset only in a more reduced fashion in the Logos, Metal (§. 3.3.2) and Eurotra (§. 3.3.1) systems. On the other hand, differently from most of the available GPCLs, in AOCLs a crucial place is occupied by collocational information, multiword expressions, and example-based multilingual correspondences, extracted from corpora and MRDs. While less directly connected to

specific theoretical frameworks, the structure and organization of AOCLs heavily reflect the needs and specificity of the systems they are part of. However, they also presents a large degree of overlapping in the adopted architecture, design and strategy.

A basic dichotomy exists in the surveyed resources, reflecting the major partition in the MT field between *interlingua-based systems* and *transfer-based systems*. Eurotra, Metal, Logos, Microsoft and Systran (§. 3.3.5) are all based on a transfer technology, and thus provide a large number of expressive devices to establish transfer conditions to cover a wide range of lexical cases and phenomena (cf. for instance the list of linguistic phenomena in §. 5.2). On the other hand, EDR is also partially based on an interlingua, which also lies at the core of the Lexical Conceptual Structure Lexicons (§. 3.3.6).

The Verbmobil lexical resources (§. 3.3.8) provide an important example of spoken lexicons, specifically geared to speech-to-speech translation. Actually, Verbmobil experience raises crucial issues for lexical resources development in general, by highlighting specific information types particularly needed by applications dealing with spoken language, and that are usually lacking in lexicons oriented to written text (e.g. phoneme patterns, enhanced with prosodic information such as syllable boundary and stress marking, pronunciation variants, lexicalised discourse phenomena such as hesitation markers, etc.). Thus, *spoken language lexicography* clearly emerges as an important extension-complementation of the more traditional and already well-established computational lexicography. What the Verbmobil experience shows is, in fact, that speech-to-speech translation systems need to access both traditional linguistic information (morphologic, syntactic and semantic), and speech-specific lexical information.

A great amount of overlapping actually exists between GPCLs and AOCLs (whose information types are a subset of those encoded in the former resources), together with also a high degree of complementarity. In fact, AOCLs in most cases lack some pieces of explicitly represented semantic knowledge, which could be employed in establishing more complex and articulated transfer conditions, while *vice versa* GPCLs are in many cases still deficient on the side of multilingual connections as well as in the encoding of corpus-based examples of language-to-language mappings. This complementarity can be extremely useful in representing an important road towards a deeper integration between those two types of resources, in the quest for a common parlance that might enhance the interchange of information and the dialogue between theoretical research and applicative needs.

6.1.4 Lexical data representation and interchange formats (LDRIF)

GENELEX (§. 3.3.8 and OLIF (§.3.3.2) represent interesting and successful examples of general models for lexical data representation and lexicon development. They have both important instantiations in concrete resources, i.e. SIMPLE/PAROLE lexicons for GENELEX and Metal and Logos lexicons for OLIF. Besides this, GENELEX also offers a wide, extensible and highly expressible language for the representation and encoding of monolingual and multilingual lexical information. The result is a relational model for lexicon organization, which assures modularity and scalability of the resources. OLIF is also particularly geared towards lexicon resource development, besides a particular attention to the representation of meta-data information, which are crucial in the process of lexicon construction, reuse and versioning. While lacking the same coverage of semantic information types as GENELEX, OLIF actually offers extremely rich expressive tools to deal with complex lexical transfer relations and transformations that occur in multilingual mappings.

It is important to stress that both GENELEX and OLIF act as interchange formats for lexical data, which allow for the development of reusable resources and parallel lexicons. While this is the natural and first vocation of OLIF, GENELEX too provide a standard representational model for the

lexicon, which is highly EAGLES compatible and guarantees data exchange and portability. Therefore, both GENELEX and OLIF surely represent important reference and starting points for the ISLE work of standard definition

6.2 A roadmap for ISLE

The purpose of the survey was to provide the necessary indication for the ISLE CLWG standardization work, as directly stemming from the state of the art in multilingual lexical resources as well as from the current needs of existing HLT systems. We can well say that this objective has been fully achieved and that the analysis of the available resources illustrated in the above sections has highlighted some *hot issues* that lie at the core of the process of defining standard for multilingual lexicons at the service of the HLT community. In this section we will illustrate some of these issues, composing the roadmap that will guide and orient the next steps of the CLWG work:

1. *Theoretical frameworks mapping and integration* – In many cases, there are resources that, although developed according to different theoretical frameworks, seem to offer fairly similar and highly compatible types of lexical information. An effort towards a more in-depth analysis of the differences and similarities between these resources, their theoretical solutions and their contents, would surely enhance the chances of data integration and exchange, as well as the portability of the resources. The issue is not framework independence, but rather to establish the proper mappings between the types of information and representations that different resources offer. In other terms, the purpose should be *to let each resource speak its own jargon, but make them understand each-other, when this is really possible*.

2. *Explicit and implicit linguistic knowledge integration* – A large scale contrast revealed by the survey is the one between linguistic information that is *explicitly represented* through some kind of representational language (i.e. ontology, conceptual structures, subcategorization frames, semantic relations, etc.), and linguistic information that is *implicitly encoded* through example patterns, collocational expressions, etc., and which is widely used in many multilingual applications. An important task is to find the way to synergically integrate both types of information in lexical resources, in order to allow systems to simultaneously access them. In fact, it seems that in order to optimally operate in truly multilingual environments, it is not possible to ignore either of these types of information.

3. *Lexical resources as distributed environments* – Lexicon construction is an highly costly enterprise, and a major goal is to set up general infrastructures to ease and optimise this process. The crescent needs of lexical data, both of general and of domain-specific nature, makes lexicon development an always incremental and potentially open effort, often to be carried out in distributed environments and through the joint work of multiple actors. It is therefore necessary to facilitate lexicon versioning and authoring, the fast integration and scalability of the resources, the fast integration of domain and general linguistic knowledge, as well as the integration of the work of human lexicographers with the information automatically extracted from corpora and dictionaries. A not very far future would in fact see the possibility to simultaneously access multiple resources, each with different types of information or more geared towards certain domains, and each developed independently or distributed on different locations and repositories.

4. *Towards multimodal resources* – The emergence of technologies like speech-to-speech translation and multimodal applications establish a new frontier for lexical resources, the one in which linguistic information traditionally encoded for written HLT is paired with the representation of information which is specifically requested for multimodal tasks. Integrated resources seem to be what systems will really need in the near future, and which would make computational lexicons truly up to the developments in HLT.

The standardization enterprise pursued by the current ISLE CLWG cannot hope to cover all these aspects, which nevertheless must form the general reference scenario for its work. Actually, some of the above points, being more firmly established and investigated, seem to offer themselves to a faster and easier standardization, while others do really belong to the still waving and uncertain frontier between advanced research and assessed technology. Thus the CLWG work must find the delicate and crucial balance of proposing a standard framework for well-established lexical solutions in multilingual environments, while being open towards the next generation of systems and their correlated needs. Actually, two final major aspects are worth stressing. First of all, the scenario of multilingual lexical resources reveal a great amount of *complementarity* among the solutions offered by existing typologies of resources. This complementarity makes integration possible and actually desirable, as one of the most expected results from the ISLE CLWG work. Secondly, standardization proposals should not lead to the elaboration of another off-the-shelf lexical architecture or formalism, but rather to the development of a *meta-scheme* for the representation, integration and exchange of lexical information in multilingual environments. Such a meta-scheme must be regarded as answer to the need of moving towards the definition of a common parlance among different actors in the HLT and among different typologies of lexical resources, so as to ensure a fair information transfer from different resources, fostering the developments and enlargements of lexical knowledge-bases, and enhancing their effective exploitation by HLT systems.

References

- Alonge A., Bertagna F., Calzolari N., Roventini A., Zampolli A. (2000), *Encoding information on adjectives in an Italian semantic net for computational applications*, NAACL-2000 Proceedings, Seattle.

- Alonge A., Calzolari N., Vossen P., Bloksma L., Castellon I., Marti M. A., Peters W., (1998), *The Linguistic Design of the EuroWordNet Database*, in Computers and the Humanities, Special Issue on EuroWordNet, Vol. 32, Nos. 2-3.

- Antoni-Lay M-H, Francopoulo G. and Zaysser L., *A Generic Model for Reusable Lexicons : The GENELEX Project*, in Linguistics and Literary Computing. vol. 8 n°4, Oxford.

- 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. and 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 Rubio A., Gallardo N., Castro R., Tejada A. (eds.), Proceedings of the First International Conference on Language Resources and Evaluation, Granada, pp.217-224.

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

- Baker, Collin F., Fillmore C. J., and Lowe John B. (1998), *The Berkeley FrameNet project*; Proceedings of the COLING-ACL, Montreal, Canada.

- *Collins-Robert English-French dictionary* (1978), 1° edition, Atkins & Duval (eds.), Harper Collins Publishers and Dictionnaires Le Robert, Glasgow and Paris.

- Corazzari O., Calzolari N., Zampolli A. (2000), *An Experiment of Lexical-Semantic Tagging of an Italian Corpus*, Proceeding of LREC-2000.

- Dorr B. J. (1993), *Machine Translation: A View from the Lexicon*, Cambridge, MA: MIT Press.

- Dorr B. J. (1997), *Large-scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring*, in Journal of Machine Translation, 12: 4, pp. 271-325.

- Dorr, B. J., Hendler J, Blanksteen S., and Migdalof B. (1995), *Use of LCS and Discourse for Intelligent Tutoring: On Beyond Syntax*, in M. Holland, J. Kaplan, and M. Sams (eds.), Intelligent Language Tutors: Balancing Theory and Technology, Lawrence Erlbaum Associates, Hillsdale, NJ, pp.289-309.

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

- EAGLES/LEXICON/MORPHOSYNTAX GROUP (1995), *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora*.

- EDR (1988) *Electronic Dictionary Project*. Japan Electronic Dictionary Research Institute Ltd. April 1990.

- EDR (1990a) *An Overview of the EDR Dictionaries*. EDR Technical Report TR-024. Japan Electronic Dictionary Research Institute Ltd. April 1990.

- EDR (1990b) *Japanese Word Dictionary*. EDR Technical Report TR-025. Japan Electronic Dictionary Research Institute Ltd. April 1990.

- EDR (1990c) *English Word Dictionary*. EDR Technical Report TR-026. Japan Electronic Dictionary Research Institute Ltd. April 1990.

- EDR (1990d) *Concept Dictionary*. EDR Technical Report TR-027. Japan Electronic Dictionary Research Institute Ltd. April 1990.

- EDR (1990e) *Bilingual Dictionary*. EDR Technical Report TR-029. Japan Electronic Dictionary Research Institute Ltd. April 1990.

- EDR (1990f) Proceedings of the International Workshop on Electronic Dictionaries. November 8-9, 1990, Oiso, Kanagawa, Japan. Japan Electronic Dictionary Research Institute Ltd.

- EDR *Dictionary Development Support System*. EDR Technical Report TR-015. Japan Electronic Dictionary Research Institute Ltd. (1989)

- Fontenelle T. (1997), *Turning a bilingual dictionary into a lexical-semantic database*, Max Niemeyer Verlag, Lexicographica Series Maior 79, Tubingen.

- GENELEX Consortium (1993), *Report on the Syntactic Layer*, Project EUREKA GENELEX, Version 4.0.

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

- GENELEX Consortium (1994b), *Report on the Morphological Layer*, Project EUREKA GENELEX, Version 3.3.

- GENELEX Consortium (1994c), *Rapport sur le multilinguisme*, Project EUREKA GENELEX, Version 2.0.

- GENELEX Consortium (1994d), *Couche morphologique- Adaptation à la langue anglaise*, Project EUREKA GENELEX, Version 1.1.

- GENELEX Consortium (1994e), *Couche syntaxique- Adaptation à la langue anglaise*, Project EUREKA GENELEX, Version 1.0.

- Gerber L.,Yang J. (1997), *SYSTRAN MT Dictionary Development*, paper presented at MT Summit V, San Diego.

- Gibbon D., Mertins I. and Moore R. eds. (2000), *Handbook of Multimodal and Spoken Dialogue Systems*, Dordrecht, Kluwer.

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

- Gonzalo J., Verdejo F., Peters C. and Calzolari N. (1998), *Applying EuroWordNet to Cross-Language Text Retrieval*, in Computers and the Humanities, Special Issue on EuroWordNet, Vol. 32, Nos. 2-3,.

- 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.

- IPAL (1988) The IPA Lexicon of the Japanese Language for Computers (Basic Verbs). Technical Note. Software Technology Center, Information-technology Promotion Agency, Tokyo, Japan.

- ISED (1988) ISED ’88. Manuscripts & Program. Proceedings of the International Symposium on Electronic Dictionaries. November 24- 25, 1988, Tokyo, Japan. Inter Group Corp (ISED ’88).

- Kakizaki N. (1987), *Research and development of an electronic dictionary: Current status and future plans*, in MT Summit Manuscripts and Program. JEIDA, Tokyo. pp. 67-64.

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

- 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.

- Levow G., Dorr B. J., Lin D. (2000), *Construction of chinese-english semantic hierarchy for cross-language retrieval*.

- Martin W. and Tops G. A. J. (eds.) (1986), Groot Woordenboek Engels-Nederlands, Van Dale Lexicografie, Utrecht.

- Miike S. (1990), *How to define concepts for electronic dictionaries*. In EDR (1990f).

- Miike S., Amano S., Uchida H. & Yokoi T. (1990), *The structure and function of the EDR Concept Dictionary*, in Czap H. & Nedobity W. (1990), TKE ’90: Terminology and Knowledge Engineering. Indeks Verlag.

- Miller G.A, Beckwidth R., Fellbaum C., Gross D., and Miller K.J. (1990), *Introduction to WordNet: An On-line Lexical Database*, in: International Journal of Lexicography, Vol 3, No.4 (1990), 235-244.

- Miwa K. (1990), *The story of demonstration session*, in EDR (1990f).
- 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.
- Muraki K. (1990), *Machine translation systems and a large scale dictionary*, in EDR (1990f).
- Nakao Y. (1990), *How to extract dictionary data from the EDR corpus*, in EDR (1990f).
- Normier B., Nossin M. (1990), *GENELEX Project: EUREKA for linguistic engineering*,Proceedings of international workshop on electronic dictionaries, Oiso, Kanagawa, Japan.
- Peters W., Vossen P., Dìez-Orzas P., Adriaens G. (1998), *Cross-linguistic Alignment of WordNets with an Inter-Lingual-Index*, in Computers and the Humanities, Special Issue on EuroWordNet, Vol. 32, Nos. 2-3.
- Pustejovsky J. (1995), *The Generative Lexicon*. Cambridge, MA, MIT Press.
- Roventini A., Alonge A., Bertagna F., Magnini B., Calzolari N. (2000), *ItalWordNet: a Large, Semantic Database for Italian*, Proceeding of LREC-2000.
- 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.
- Sanfilippo A. *et al* (1999), *EAGLES Recommendations on Semantic Encoding.* See http://www.ilc.pi.cnr.it/EAGLES96/rep2
- Sanfilippo A. *et al.* (1996), *EAGLES Subcategorization Standards*. See http://www.ilc.pi.cnr.it/EAGLES96/syntax/syntax.html.
- Sanfilippo A. *et al*. (1995), EAGLES /LEXICON/SYNTAX GROUP, *Report on the syntax.*
- Thurmair G. (1990), *Complex Lexical Transfer in METAL*, in Proceeding TMI.
- Thurmair G. (2000), *OLIF Input Document*. See http://www.olif.net/main.htm.
- Tsurumaru H. (1990), *Extraction of semantic hierarchical relations from a Japanese language dictionary*, in EDR (1990f).
- Uchida H. (1988), *Towards common knowledge for natural language processing*, in ISED (1988), pp. 92-93.
- Uchida H. (1990), *The EDR Electronic Dictionaries*, in EDR (1990f).
- Underwood N. and Navarretta C. (1997), *A Draft Manual for the Validation of Lexica*. Final ELRA Report, Copenhagen.
- Villegas M. *et al*. (2000), *Multilingual linguistic resources: from monolingual lexicons to bilingual interrelated lexicons*., LREC-2000.

- Vossen P. (1998), *Introduction to EuroWordNet*, in Computers and the Humanities, Special Issue on EuroWordNet, Vol. 32, Nos. 2-3.

- Vossen P., Bloksma L., Peters W., Kunze C., Wagner A., Pala K., Vider K. (2000), *Extending the Inter-Lingual-Index with new concepts*, Deliverable 2D010, EuroWordNet, LE2-4003.

- Wahlster W. (ed.) (2000), *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin: Springer.

- Yokoi T. (1988), Electronic dictionary development in Japan. In ISED (1988), pp. 97-100.

- Yokoi T. (1990), *Collaboration and cooperation for development of electronic dictionaries – Case of the EDR Electronic Dictionary Project*, in EDR (1990f).

- Yokoi T., Uchida H., Amano S. and Kiyono (1989), *M. Research and development of large-scale electronic dictionaries*, Proceedings of the First Australia-Japan Joint Symposium on Natural Language Processing.

- Yokota E. (1990), *How to organize the concept hierarchy*, in EDR (1990f).

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