Towards a Standard for the Creation of Lexica

May 2003

Monica Monachini*
Francesca Bertagna*
Nicoletta Calzolari*
Nancy Underwood**
Costanza Navarretta**

* Istituto di Linguistica Computazionale - CNR
Via Moruzzi 1
56100 Pisa
Italy

**Center for Sprogteknologi
Njalsgade, 80
DK 2300 Copenhagen S
Denmark

Acknowledgements:
We would like to thank the experts from ELRA Panel for Validation of Written Language Resources (EPV-WLR) who gave valuable feedback on earlier versions of this document.

Contents

1. Introduction

This document has been developed for ELDA in two distinct phases.

The first version was born in 1997 in tandem with the Draft Manual for the Validation of Lexica (Underwood & Navarretta 1997). Being firmly based on the work carried out by the "first" EAGLES and reported in (Monachini & Calzolari 1996 and Sanfilippo *et al.*1996), it constituted a draft proposal for a standard for the creation of computational lexica, with a view also to aiding the process of validating lexica. It focussed on Morphosyntax and Subcategorization. The much more detailed specifications for Italian, French and German (along with draft specifications for English) developed by the EAGLES members served as complement to the first version of this document[1], especially as far as morphosyntax is concerned.

The present version has been developed when the EAGLES recommendations were worked out for Lexical Semantics as well, but especially after (i) the PAROLE-SIMPLE and EuroWordNet experiences and (ii) the ISLE project (the "second" EAGLES). Within PAROLE-SIMPLE[2], the EAGLES recommendations for subcategorization and lexical semantics were concretely applied, revised and re-elaborated in view of the creation of plurilingual lexica. During ISLE, all the EAGLES bulk of work was exploited and its results extended in a multilingual perspective, trying to make a synthesis of all the information relevant to build a multilingual lexical entry (a MILE) starting from a monolingual description.

The aim of this document is, hence, to provide an analysis of the so-called *basic notions*, i.e. linguistic information crucial for (i) the description of a computational lexical entry and (ii) lexicon validation as well, from a monolingual point of view at the morphosyntactic, syntactic and semantic levels. Then, all the notions needed for going from a monolingual to a multilingual entry are presented.

The main input to this work comes from the previous experiences, i.e.:

- the Recommendations on Morphosyntax (Monachini and Calzolari 1996, available for browsing and download at http://www.ilc.pi.cnr.it/EAGLES96/morphsyn.html) for the morphosyntactic level.
- the Recommendations on Subcategorization (Sanfilippo *et al.* 1996, available for browsing and download at http://www.ilc.pi.cnr.it/EAGLES96/synlex/synlex.html) for the syntactic level.
- the Recommendations on Lexical Semantics (Sanfilippo *et al.* 1999 available at http://www.ilc.pi.cnr.it/EAGLES96/EAGLESLE.PDF), where already emerged a very large set of agreed-on information.
- the syntactic and semantic layers of the PAROLE and SIMPLE lexicons (www.gilcub.es)[3]. They, built-up with the flexible and harmonized GENELEX model, uniform criteria and types of information for twelve EU languages, can be seen as plurilingual lexicons (Lenci *et al.* 2000a).
- the ISLE Survey of main approaches towards bilingual and monolingual lexicons (Calzolari *et al.* 2001), which provides an examination of linguistic phenomena crucial to sense distinction and to the selection of the correct translation equivalent.

[1] Those ELM specifications can be found at the EAGLES website (http://www.ilc.pi.cnr.it/EAGLES/home.html) and answer some of the difficulties mentioned in the section devoted to morphosyntax.
[2] – and also in their national extensions (e.g. CLIPS, Ruimy et al. 2002) –
[3] Cf. Ruimy *et al.* forthcoming, for the SIMPLE Italian Lexicon.

– the ISLE Deliverable carried out by the Computational Lexicon Working Group, where a compendium of the basic notions crucial to the creation of a lexical entry and the operations on those needed to arrive at a multilingual lexical entry (the MILE) is provided (Calzolari *et al.* 2002)[4].

2. A proposal for a standard

The next sections (from 3 to 6) are built on the achievements of the EAGLES specifications by attempting to integrate them where needed. We propose that, in future, this EAGLES-based standard could be used in the validation of lexica. In addition, we foresee that the proposal here also function as an aide memoir or checklist for validators in designing their language specific part of the validation (see e.g. Underwood & Navarretta 1997).

In section 3 we describe the morphosyntactic notions[5] which should possibly be included in a lexicon and also how we have generalised somewhat from the information presented in EAGLES (Monachini and Calzolari, 1996).

In section 4, we describe the notions needed for encoding subcategorisation information, from EAGLES (Sanfilippo *et al.* 1996), from the PAROLE instantiatiations and the recent ISLE experience (Calzolari *et al.* 2002).

In section 5 the information needed to describe the semantic level are presented, from EAGLES, from the SIMPLE experience (Lenci et al. 2000b) and ISLE (Calzolari *et al.* 2002).

Finally, section 6, starting from the ISLE focus on multilingual (Calzolari *et al.* 2002), provides the set of operations to be performed on the monolingual notions for building the multilingual level.

A general presentation of the basic notions for each level of description, i.e. information concurring to define e.g. a morphosyntactic unit, a syntactic structure, a semantic predicate or a multilingual correspondence will be provided by means of examples. These, when needed, will be also described in terms of their constitutive sub-elements.

Following the ISLE approach (Calzolari *et al.* 2002), we aim "to reach a maximal decomposition into the minimal basic information units that reflect the phenomena we are dealing with". This principle is used to allow easier reusability or mappability into different theoretical or system approaches (Heid and McNaught 1991): small units can be assembled, in different frameworks, according to different (theory/application dependent) generalization principles. Lexica are built for different purposes and users and can be specialised so that they only cover a few linguistic phenomena (valency, linear order etc.), only describe one category (verbs, nouns etc.) or apply to specific NLP systems and/or applications. All these differences can have a repercussion on features more or less important in a lexicon. The basic notions

[4] The document is the outcome of a strong collaboration within a group of experts constituted by European, American and Asian partners: Sue Atkins, Nuria Bel, Francesca Bertagna, Pierrette Bouillon, Nicoletta Calzolari, Thatsanee Charoenporn, Dafydd Gibbon, Ralph Grishman, Chu-Ren Huang, Asanee Kawtrakul, Nancy Ide, Hae-Yun Lee, Alessandro Lenci, Paul J. K. Li, Jock McNaught, Monica Monachini, Jan Odijk, Martha Palmer, Valeria Quochi, Ruth Reeves, Dipti Misra Sharma, Virach Sornlertlamvanich, Tokunaga Takenobu, Gregor Thurmair, Marta Villegas, Antonio Zampolli, Elizabeth Zeiton.

[5] We chose to comply with the ISLE terminology and call all the information necessary at each level of linguistic description *basic notion*.

must be established before considering any system-specific instantiation, otherwise their finding may be too conditioned by system-specific approaches.
For example, 'synonymy' can be taken as a basic notion; however, the notion of 'synset' is a generalization(specialization), closely associated with the WordNet approach. 'Qualia relations' are another example of a generalization(specialization), whereas 'semantic relation' is a basic notion. Modularity is also a means to achieve better granularity. High granularity and maximal decomposition does not mean that we limit our recommendations to these very basic notions. On the contrary, whenever consensus has been found on a more complex linguistic object, we provide such shareable commonly agreed linguistic objects (e.g. synsets and qualia relations).

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

The lexical notions are formalized by means of the *MILE Lexical Classes* (MLC), that represent the main building blocks of the lexical entries. The MILE model provides the definition of these classes, i.e. their attributes and the way they relate to each other (some complex classes are defined in terms of other classes). Classes represent notions like *syntactic feature*, *syntactic phrase*, *predicate*, *semantic relation*, *synset*, etc. The instances of MLC are the *MILE Data Categories* (MDC). So for instance, NP and VP are data category instances of the class <Phrase>, and SUBJ and OBJ are data category instances of the class <Function>. Each MDC is identified by a URI. MDC can be either "user defined" or belong to "shared repositories".

3. Morphosyntactic Information

The EAGLES specifications for Morphosyntax are the result of a bottom-up approach, consisting of a comparison of the main encoding practices in lexica and corpora and resulting in a consensual proposal on the basis of this comparison. The consensual proposal has been tested by applying it to Catalan, Danish, Dutch, French, English, German, Greek, Irish, Italian, Portuguese, Spanish, Swedish. This testing phase was based on existing lexica and, for some languages, tagged corpora in the respective languages.

These recommendations – by the way they came into existence, i.e. based on commonly accepted practices – constitute a detailed agreed on platform which a block of lexicons, e.g. the PAROLE lexicons but also many others, are built on. Firstly developed starting from the perspective of the languages of the EU Community, later, they have been extended also to cover the requirements and peculiarities of Eastern EU languages in the framework of the MULTEXT-East Copernicus Project. The outcomes of this experiment are reported in Monachini 1995 and Erjavec & Monachini 1997 that can be seen as complement to this document for the construction (and validation) of morphosytactic lexicons for Eastern languages[6].

Making a consensual proposal based on a variety of languages means that even in such a thorough approach as EAGLES, there is still room for different interpretations within the "standard". Certain sorts of information can be arranged in various ways, possibly without detriment to the value of the lexicon to particular users. For example whilst for the major word classes (verb, noun) the category is generally agreed upon across languages and lexica, this is not the case with minor word classes, for example, the division of categories into determiners and articles could be collapsed into one category. Similarly the classification of certain word classes such as possessives differs from language to language, so that they may be a type of pronoun or adjective or determiner. Various differences in category assignment can either be due to the nature of the language itself or because of different lexicographic traditions associated with different languages.

The EAGLES language independent morphosyntactic specifications have been divided into three levels (1) obligatory (grammatical categories), (2) recommended (a minimal common core set of features), and (3) optional (information not usually encoded in more than three languages or not purely morphosyntactic). Note that this means that certain so-called "optional" information is actually to be strongly recommended for those languages to which it applies.

In the following we take each word class in turn and present the language independent specifications in the same way as in EAGLES, followed by glossary and explanation. Then these are applied to three specific languages (Danish, English and Italian) to indicate how the specification can apply to a specific language. In an attempt to make these specifications more generally applicable to a wider range of lexica, we have in some cases made some generalisations over the EAGLES proposals.

[6] The EAGLES work covers a wide range of Indo-European languages which are found in Europe. Once non Indo-European languages are included it is clear that many EAGLES features could not be applicable, on the other hand many features relevant for such languages are necessarily missing from the EAGLES specifications. Therefore the only information which is really *obligatory* for all languages is the Category information, and only when the same Category is recognised in all languages for the same word class. When this is not the case it must be specified for each language how the given Category is related to the proposed specifications.

In addition, at the language independent level we have added an extra feature. Because the EAGLES specifications have been built up on the basis of both lexica and tagsets for corpora, the features reflect fully inflected forms. It is, of course, not the case that all lexica, (even rich ones) would contain fully inflected forms. Therefore we have added an extra attribute to account for inflectional patterns and/or irregular forms in those cases where the morphosyntactic features are defined via inflection and so may not be present in e.g. a stem dictionary.

When values are given in parenthesis, (), it means that they are language dependent e.g. in the case of Gender we put the values *masculine, feminine, common, neuter, generic* in parenthesis (this notational distinction is not in the EAGLES specifications). A feature marked with a star (*) indicates that the given feature may be inflectionally realised and in that case it would only apply to full-form entries (e.g. Number is applicable to full-forms, Gender otherwise pertain to the lemma level as well). A number of such features, although generally deriving from inflection may sometimes be inherent in the lexical item, e.g. in Danish definiteness on a noun may be realised inflectionally, however proper names by definition are inherently definite.

In Monachini & Calzolari 1996 the gender *common* is given as language specific for Italian and Spanish. It refers to the cases where it is impossible to decide whether a lexical item can refer to something which is either feminine or masculine (e.g. names of professions where the person referred to could be either male or female) and it is only within context that the gender can be determined. However, in Danish, *common* is one of the two possible genders (the other one being *neuter*). Thus this value in Danish is rather different from *common* in Italian and Spanish. Therefore we have decided to re-name the EAGLES value *common, generic* (gn) as a super-category for feminine and masculine.

In the following we present the recommended features for each word class and the specific features for Danish, English and Italian. In the language independent tables for each class, the rows numbered 1, 2, and 3 reflect the three levels (obligatory, recommended and optional) introduced in EAGLES. In some cases, no value is given for a feature in the general table, this is because that feature is specific to a language not specifically treated here.

Noun

	Cat	Type	Gend	Numb*	Case*	Count	Defin*	Noun Infl
1	Noun							
2		com prop	(m/f/ c/n/gn)	(sg/pl inv)				
3					(nom gen/dat/ acc)	coun mass		

The Category **Noun** is common to all languages, and is obligatory.
Recommended information is:

- **Type** common, proper.
- **Gender** Depending on the language the values are: masculine, feminine, common, neuter and generic. The generic Gender value is a suggestion for expressing a Boolean disjunction in cases where a noun can refer to objects with either masculine or feminine gender.
- **Number** singular, plural, invariant. The invariant Number value also expresses a Boolean disjunction in that the singular and plural forms for a noun are the same.

Optional information is:

- **Case** Values are: nominative, genitive, dative and accusative when this is relevant to a particular language.
- **Countability** mass, count (coun).
- **Definiteness** applies to the Scandinavian languages (enclitic definite articles).
- **Inflection** (Noun Infl) was originally presented as a Danish/German feature in EAGLES, but can be used to give the inflectional type in many languages, in particular when the lexicon is not full-form. In addition this can contain information on irregular or unpredictable inflectional forms (e.g. En: man, men)

Number, Case and Definiteness are marked with a star because they are, in most cases, inflectional features. Irregular or unpredictable inflectional forms should be given also in lexica which are not full-form i.e. as a value for Noun Infl.

An attribute called Declination was included in the original EAGLES proposals to account for German noun declensions, however it seems that this could be covered by the Infl feature, and so it has been omitted.

Information for Danish Nouns

	Cat	Type	Gend	Numb*	Case*	Count	Defin*
1	Noun						
2		com prop	c n	sg pl			def unmk
3					gen no-gen	coun mass	

The valid values for Gender in Danish are common and neuter. Case values are genitive and non-genitive. The Definiteness (defin) attribute marks the presence (def) or absence (unmk) of the enclitic article.

Information for English Nouns

	Cat	Type	Numb*	Count
1	Noun			
		com prop	sg pl	
				coun mass

There is no Gender distinction for English, thus the relevant attributes are Category, Type and Number.

We have also added the attribute Count since although this is not a morphological distinction, and was not included in the EAGLES application, it is a syntactic distinction which does apply to English.

It must be noted that the Case attribute may or may not be applied to English Nouns, depending on whether the clitic *'s* is considered a marker of the genitive case for nouns or a postposition. We have chosen the second solution because it is the one proposed in the EAGLES English application, but both solutions are fully acceptable.

Information for Italian Nouns

	Cat	Type	Gend	Numb*	Count
1	Noun				
2		com prop	m f gn	sg pl inv	
3					coun mass

In Italian, it is possible to give a value generic for Gender when a noun can be both masculine and feminine (e.g. *dentista*, dentist). The Number value invariant can be used for encoding invariant nouns (e.g. *città*, town/towns). Case is not a feature pertinent to Italian.

VERB

	Cat	Type	Fin*	Vf-M*	Tens*	P*	N*	G*	Asp*	Vce*	Refl	Auf	Ax	Se	Clt	Zu	Verb Infl
1	Verb																
2		(main no-main)	fin no-fin	(ind subj imp con infi part ger sup ing-frm)	(pres impf fut past)	(1 2 3)	(s p)	(m f c n)									
3		s-au cop							pfve ifve	act pas	refl no-refl	prg prf pss pph					

The Category **Verb** is recognised in all languages, and is obligatory.

Recommended information is:

- **Type** main or non-main (no-main).
  (The verbal types suggested by EAGLES are main, auxiliary, and modal, but different languages/theories do not agree with this distinction. We suggest as general types main and non-main. Modal could then be a subtype of the type main or non-main, according to the modal characteristics in the different languages).

- Finiteness (**Fin**), indicates whether the verb is finite (fin) or not (no-fin).

- **Vf-M** this attribute collapses the two notions of Verb-form and Mood together. The features Finiteness, Verb-form and Mood can be coded differently in Germanic and Romance languages depending on different traditions for how the distribution of finite and non finite verb forms is described. See the language specific applications for the different ways these can be split up and the dependency between Finiteness and Verb-form/Mood features.

- **Tense**, (T) has the possible values: present (pres), imperfect (impf), future (fut) and past.

- **Person** (P), **Gender** (G) and **Number** (N).

Optional features are:

- **Aspect** (Asp) with the values: perfective (pfve) and imperfective (ifve).
- **Voice** with the values: passive(pas) and active (act).
- **Reflexivity** (Refl) with the values: refl and no-refl.
- **Auxiliary Function**, indicating the function of auxiliaries: progressive (prg), perfect (prf), passive (pss), and periphrastic (pph).
- **Verb Infl** can be used to encode either inflectional patterns and/or irregular forms.

Finiteness, Verb-Form/Mood, Tense, Person, Number, Gender, Aspect and Voice are often formed via inflection, thus they have been marked with a star.

The original EAGLES proposal also had the feature Main-Verb function whose values were transitive, intransitive or impersonal. However, this appears to be superseded by the use of the attribute Frame and so it has been left out.

Some of the language specific features which can be encoded:

- **Auxiliary** (Ax) encodes information concerning the choice of auxiliary for compound tenses.
- **Separability** (Se) is used in Dutch for verbs with separable particles.
- **Clitic** (Cl) indicates, in some languages, the presence/absence of a clitic.
- **Zu** feature for German indicates the infinitive incorporating "zu".

Information for Danish Verbs

	Cat	Type	Fin*	Vform*	Mood*	Tens*	Voice*	Aux-Type
1	Verb							
2		main no-main	fin no-fin	 infin past-part pr-part	indic imper	pres past	non-s-pass s-pass	ax-act ax-pass

In Danish, the two Types main, and non-main are recognised. A subtype of main is "modal".

Contrary to the EAGLES guidelines, the features verb form (Vform) and Mood have been separated. The Finiteness feature may be superfluous in that there is a strict dependency so that Vform values are non-finite (infinitive, perfect participle, present participle) whilst Mood values are finite (indicative, imperative). However, the higher level finite/non-finite distinction is often useful in processing.

The two values for Tense are present and past.

The Voice feature is language specific distinguishing between the *s*-passive form (s-pass) and all other forms (non-s-pass) .

The feature Auxiliary Type distinguishes among auxiliaries used to form compound tenses (ax-act) and the auxiliary *blive* (ax-pass) which combines with a past participle to form the passive.

Danish verbs do not inflect for person, number or gender.

Information for English Verbs

	Cat	Type	(Fin*)	V-form*	Mood*	Tense*	P*	N*	Aux-Type
1	Verb								
2		main no-main	fin no-fin	 infinite ing-form particip	indic subj imper	pres past	1 2 3	sg pl	primary modal

As with Danish, the Vform and Mood features have been separated and the feature Finiteness may be superfluous because V-form values (infinitive, ing-form and past participle), depend on the verb being non-finite and Mood values (indicative, subjunctive and imperative) depend on it being finite.

The Tense attribute has the two values present and past.

English verbs also inflect for Person and Number.

The language specific feature Auxiliary Type is introduced distinguishing between primary auxiliaries (be, have) and modals.

Information for Italian Verbs

	Cat	Type	Finite*	V-fM*	Tens*	Pers*	Numb*	Gend*	Clt*
1	Verb								
2		main no-main	finite no-finite	indic subj imper cond infin part gerund	pres imperf futur past	1 2 3	sg pl	masc fem	clitic no-clitic

Following the EAGLES guidelines and in contrast to Danish and English, the V-fM features are not separated for Italian. However, there is still a dependency between Finiteness values and V-fM features, so that finite verbs have the possible V-fM values: indicative, subjunctive, imperative, conditional and not-finite ones have the possible V-fM values: infinitive, participle, gerund).

The Tense values are present, imperfect, future and past.

Italian verbs inflect for Person, Number and Gender.

The language specific feature Clitic refers to the pronominal particles which can accompany verbs in order to make pronominal, reflexive or reciprocal forms.

ADJECTIVE

	Cat	Type	Degr*	Gend*	Numb*	Case*	Use	Mod	Adj Infl	Pos	Pers	Defin*	Comp	Frame
1	Adjective													
2		(qualif posse indef cardin ordin)	posit compar super	(m f n c gn)	(sg pl)								(infl peri)	
3						(nom gen dat acc)	(attrib predic adverb nomin)	premod postm		sg pl	1 2 3			

The Category **Adjective** is obligatory.

Recommended information is:

- **Type** (suggested values are qualificative, possessive, ordinal, cardinal and indefinite). This range of values allows for different category assignments in different languages. Qualificative apparently applies to the core set of 'normal' adjectives on which there is general agreement. However certain Romance languages (e.g. French) classify possessives as a type of adjective rather than as pronouns or determiners. Numerals (cardinals and ordinals) could also be considered as a separate category (see the section on numerals below).

- **Degree** (positive, comparative and superlative), extra values may be necessary for some languages. Degree only applies to qualificative adjectives.

- **Gender** and **Number** are also recommended for those languages whose adjectives inflect for those features.

- **Comparison** (Comp) indicates whether the adjective inflects (infl) for degree or uses periphrastic constructions (peri). In languages where both synthetic and analytic degree form are possible, lexica should indicate which form applies to which adjective. The feature Flection (Flect) was originally introduced for German to account for this but we have replaced that feature with Comp. It is probably applicable to most languages

Optional information is:

- **Use** (most common values are attributive and predicative, but other adjectival uses such as adverbial and nominal may be given at the language specific level).

- **Modification** (Mod) indicates whether an adjective precedes (premod) or follows (postm) the noun. The default value differs from language to language.

- **Pos** and **Pers** both depend on the analysis of possessives as a type of adjective. For a possessive adjective Pos indicates the number of the possessor and Pers indicates the person of the possessor.

- **Case** is clearly only applicable to those languages which have case assignment.

- **Inflection** (Adj Infl) allows for the coding of inflectional patterns and/or exceptions where the lexicon is not full form.

The value 'normal' was included under Type and for Danish but this seems to be the same as 'qualif', and so it has been omitted from this standard.

Information for Danish Adjectives

	Cat	Type	Degree*	Gend*	Num*	Use	Defin*
1	Adjective						
2		qualif cardinal ordinal	posit compar super	c n	sg pl		defin indef unmk
3			aller-sup			attrib predic adverb nomin	

Type (qualificative, ordinal, cardinal), Degree, Gender, Number, Use and Definiteness are recognised. Here there are three possible values for Defin since adjectives can be unmarked or indicate either definiteness and indefiniteness (cf. the definiteness values for nouns).
The aller-superlative type is specific to Danish and it is formed by adding the prefix *aller-* to the superlative form to make it even stronger (*det allerbedste* (the best of the best)).

Information for English Adjectives

	Cat	Type	Degree	Mod	Use
1	Adjective				
2		qualif ordinal cardinal	posit compar super		
3				premod postm	attrib predic

To the attributes Type (qualificative, cardinal, ordinal) and Degree in the English application, we have also added Use and Mod as optional information since these also seem to be applicable.

Information for Italian Adjectives

	Cat	Type	Degree	Gend*	Num*	Use
1	Adjective					
2		qualif deter	posit comp super	m f gn	sg pl inv	
3						attrib predic

Only the two Types (qualificative and determinative) are recognised in the Italian application. Determinative adjectives include the so-called pronominal adjectives which do not take Degree, for example possessives which are syntactically adjectives e.g. *il mio libro* (lit: the my book). Degree, Gender, Number and Use (attributive and predicative) are features pertinent to Italian Adjectives.

Pronouns, Determiners, Articles

In some applications/languages the word classes Pronoun, Determiner and Article are treated as a unique class. In (Monachini & Calzolari 1996) it is proposed to distinguish three separate categories, but this is only a recommendation. Particular lexica/applications can collapse two or all three classes. As with the classification of possessives as adjectives, we cannot prescribe the category such words are assigned to but rather require that in a comprehensive lexicon, all these word types must be treated somewhere.

Pronoun

	Cat	Type	Pers	Gen	Num	Case*	Pos	Pol	Funct	Pron Infl
1	Pronoun									
2		(dem indf poss interr rela pers refl recp exc)	1 2 3	(m f c n gn)	sg pl inv	(nom gen dat acc obl pobj)	sg pl		(nom attrib pred adv)	
3								(pol fam)		

The Category **Pronoun** is obligatory,

Recommended information is:

- **Type** The suggested values are: demonstrative, indefinite, possessive, interrogative, relative, personal, reflexive, reciprocal, and exclamatory. However, the types of pronouns can vary greatly depending on whether articles and determiners are included in the category Pronoun or not, and whether certain items such as possessives are treated as adjectives.
- **Person**, **Gender**, **Case** and **Pos** (i.e. the number of the possessor) are not applicable to all pronominal Types and languages. This kind of information must be specified for each language. See the table below for the dependencies between different pronoun types and specific features in English.

Optional information is:

- **Politeness** (Pol) is relevant for personal pronouns in many languages and takes the two value familiar and polite.
- **Inflection** (Infl) was specifically introduced for French and German but again it can be applicable to all languages.
- **Function** (Funct) indicates nominative, attributive, predicative or adverbial use of a pronoun.

The feature Wh-Type was also included in EAGLES to distinguish interrogative, relative and exclamatory pronouns from other types of pronouns. However, there seemed to be some inconsistencies in its use and it has been taken out and the different wh-pronouns are treated as simple values of Type.

Dependencies between pronoun type and specific features

To help clarify the dependencies between different types of pronouns and the features which apply to them, the following table shows the dependencies for English. For a given type of pronoun, the symbol X, indicates whether a particular feature could be relevant for that pronoun type. Note that this is just an example and that such dependencies should be worked out for other languages.

Pronoun Type	Pers	Gen	Num	Case	Pos	Pol	Pron Infl
pers	X	X	X	X			
refl	X	X	X				
poss	X	X			X		
dem			X				
indf			X				
inter							
rela							
exc							

Information for Danish Pronouns

	Cat	Type	Pers	Gen	Num	Case*	Pos	Pol	Funct
1	Pronoun								
2		pers demo indf poss rela refl recp rela inter	1 2 3	m f c n	sg pl	nom gen obl	sg pl		
3								fam pol	nom attr prd adv

The recognised Types for Danish Pronouns are personal, reflexive, demonstrative, indefinite, possessive, relative, reciprocal and interrogative and relative.

The feature Gender applies to personal pronouns (only 3rd person singular), possessive pronouns and the relative interrogative "hvilken".. The Gender values feminine and masculine are only recognised for personal and possessive pronouns in third person singular. An extra feature Sexus could be introduced to hold these two values instead.

Demonstrative, personal, possessive and reflexive pronouns inflect for Number.

The feature Case applies differently to personal pronouns and to other pronouns. Personal pronouns have nominative and oblique Case values. The genitive Case occurs in possessive pronouns. Other pronouns have only genitive or non-genitive Case values.

The Politeness value polite applies to the personal pronoun *De*.

Information for English Pronouns

	Cat	Type	Pers	Gen	Num	Case*
1	Pronoun					
2		dem indf poss pers refl rela interr exl	1 2 3	m f n	sg pl	nom obl

The relevant Types for English Pronouns are demonstrative, indefinite, possessive, personal, reflexive, relative, interrogative and exclamative.

Person, Gender, Number and Case also apply to English pronouns.

Information for Italian Pronouns

	Cat	Type	Pers	Gen	Num	Case*	Pos	Pol
1	Pronoun							
2		dem indf poss pers refl inter relat exc	1 2 3	m f	sg pl inv	nom gen dat acc obj	sg pl	
3								pol fam

Italian pronouns can be divided into demonstrative, possessive, indefinite, personal, reflexive, interrogative, relative and exclamative.

Personal pronouns are inflected for Person and Number and have different Politeness values.

Reflexive pronouns inflect for Person and Number.

Possessive pronouns are inflected for Number and Gender and they agree with the nouns which they combine with.

Demonstrative, indefinite, interrogatives and exclamative pronouns inflect for gender and number. Relative pronouns also inflect for Gender and Number, with the exception of the relative “che” which does not inflect at all.

Determiner

	Cat	Type	Pers	Gen*	Num	Case*	Pos	Infl
1	Determiner							
2		(dem inter indf poss card rela exc part)	1 2 3	(m f n c gn)	sg pl		sg pl	
3						(nom gen dat acc)		

The Category **Determiner** is obligatory

Recommended information is:

- **Type** the suggested possible values are demonstrative, interrogative, indefinite, possessive, cardinal, relative, exclamatory and partitive. Again different languages and lexica may assign some of the types to different categories. What in English (and in many other languages) is called *Determiner* is in the Romance tradition classified as *Pronominal Adjective.* Pronominal Adjectives do not always correspond to Determiners (e.g. in most cases the Italian possessives are not determiners, but adjectives).

- **Person, Gender, Number** and **Pos** are also recommended depending on the type of the determiner. As with pronouns the assignment of certain features depends on the type of determiner and an example of the dependencies for English are given below.

Optional information is:

- **Case**
- **Infl**

In some cases Gender and Number can be inflectional features.

The feature Wh-Type was also included in EAGLES to distinguish interrogative, relative and exclamatory determiners from other types of determiners. However, there seemed to be some inconsistencies in its use and it has been taken out and the different wh-determiners are treated as simple values of Type.

Dependencies between determiner type and specific features

In the following table for English, for a given type of determiner, the symbol X, indicates whether a particular feature could be relevant for that determiner type Note that this is just an example for one language (English). Similar dependencies have to be worked out for other languages.

Pronoun Type	Pers	Gen	Num	Case	Pos	Infl
poss	X	X			X	
dem			X			
indf			X			
inter						

Information for Danish Determiners

	Cat	Type	Pers	Gen	Num	Pos
1	Determiner					
2		dem poss quant ordin card	1 2 3	c n	sg pl	sg pl

The Category Determiner covers lexical items which in the Danish tradition are classified as pronouns and quantifiers. Danish determiners include demonstratives, possessives, quantifiers, ordinals, cardinals.

In addition we have added the attribute Pos to account for those possessive determiners which indicate features of the possessor.

Information for English Determiners

	Cat	Type	Pers	Gen	Num	Pos
1	Determiner					
2		poss dem indef inter	1 2 3	m f n	sg pl	
3						sg pl

English determiners include possessives, demonstratives, indefinites and interrogatives.

The distinctions Person, Gender and Number apply to some determiners. Person applies to possessive determiners, while gender applies to third person singular possessive determiners.

Demonstrative determiners and some indefinite determiners (*this, much*) are inflected for Number.

The feature Pos is given as an optional rather than a recommended feature because it applies to all possessives pronouns and thus it is not inflectional in nature.

Information for Italian Determiners

	Cat	Type	Pers	Gen	Num	Pos
1	Determiner					
2		dem indf inter rela exc	1 2 3	m f gn	sg pl	sg pl

In Italian, Determiners are distinguished according to the feature Type (demonstrative, indefinite, interrogative, relative and exclamative).

Possessive pronouns are used as determiners only in combination with few family nouns, in singular form (e.g. *mio padre,* my father, *mia madre,* my mother), thus these items are not encoded as determiners.

Demonstrative, indefinite and interrogative determiners are inflected for gender and number. Indefinite demonstratives cover the class of quantifiers.

Some interrogatives (*che,* what, *quale,* which *quanto* how much) can also have exclamatory value. The relative demonstrative *il quale* is inflected for Gender and Number.

Article

	Cat	Type	Gend*	Num*	Case*
1	Article				
2		(defin indef partit)	(m f n c)	sg pl	
3					(nom gen dat acc)

In most lexica articles are treated as an independent category, but in some languages they can be incorporated in the class of Determiners or in that of Pronouns. The most common Article Types are definite and indefinite. The partitive Type has been introduced for French. Gender and Number are also recommended features, and they are often inflectional. The case feature is only relevant for some languages.

Information for Danish Articles

	Cat	Type	Gend*	Num*
1	Article			
2		defin indef	n c	sg pl

Type, Gender and Number apply to Danish articles.

Information for English Articles

	Cat	Type	Num*
1	Article		
2		defin indef	sg pl

Only Type and Number apply to English articles (*the* and *a/an*).

Information for Italian Articles

	Cat	Type	Gend*	Num*
1	Article			
2		defin indef	m f	sg pl

Type, Gender and Number are relevant features for Italian articles.

Adverb

	Cat	Type	Degree*	Polarity	Wh-T	Adv Infl	Comp
1	Adverb						
2		(general particle)	positive comparat superla				
3				wh no-wh			

The category **Adverb** is obligatory.
Recommended information is:

- **Type** (general and particle). The Type distinction between general and particle adverbs is not relevant to all languages. In some lexica, particles could be considered as case-marking prepositions or as part of a verb. In many lexica adverbs are distinguished according to their semantic value (manner, distribution, place etc.), but this is not really morphosyntactic information.
- **Degree** (positive, comparative and superlative).

Optional information is:

- **Polarity** Some languages distinguish between interrogative and non-interrogative adverbs and this information is given in the feature Polarity.
- **Wh-Type** is dependent on the adverb being interrogative and provides information on the type of interrogative
- **Adv Infl** can be used to contain inflectional paradigms or irregular forms.
- **Comp** is used to indicate whether the comparative forms of the adverb are formed periphrastically or via inflection.

Information for Danish Adverbs

	Cat	Type	Degree*
1	Adverb		
2		general particle	positive compar superla

The Type (general and particle) and the Degree distinctions are pertinent to Danish adverbs, Polarity might also be distinguished.

Information for English Adverbs

	Cat	Type	Degree*	Polarity	Wh-T
1	Adverb				
2		general particle	positive compar superla		
3				wh no-wh	rela interr excl

For English adverbs two Types can be distinguished: general and particle. The Degree, Polarity and Wh-T features (relative, interrogative and exclamative) are also relevant.

Information for Italian Adverbs

	Cat	Degree*
1	Adverb	
2		positive comparat superla

Degree is the only feature given for Italian (although Polarity could be distinguished). Adverbs can also be distinguished in Types according to their semantic value.

Adposition

	Cat	Type	Formation*	Gender*	Numb*	Case*	Ad Infl
1	Adposition						
2		(preposit postposit circumpo)					
3			simple fused	(m f n)	sg pl	nom gen dat acc	

The Category **Adposition** is a rather unusual term and it could be envisaged that the category assigned in most lexica would be one of the Type labels recommended instead, most typically preposition and postposition.

Recommended information is:

- **Type** (preposition, postposition and circumposition)

Optional information is:

- **Formation** accounts for the fact that in some languages, such as Italian and German, some prepositions can appear fused with the articles.
- **Gender**, **Number** and **Case** refer to the gender, number and, for some languages, case of the articles fused with the Adpositions.
- **Ad Infl** applies only to the inflection on fused adpositions

In some lexica, particles may be included under the Adposition Category.

Information for Danish Adpositions

	Cat	Type
1	Adposition	
2		preposition circumposition

The two Types preposition, and circumposition are recognised in Danish. Some lexica can also recognise a subtype for multi-words prepositions such as *inden for* (inside).

Information for English Adpositions

	Cat	Type
1	Adposition	
2		preposition postposition

Only the two types preposition and postposition are recognised for English. The only postposition is the genitive clitic *'s* As noted in the section on nouns the clitic *'s* can also be considered a marker for the Genitive case, in which case it will not considered as an Adposition, but as feature of the Noun Category.

Information for Italian Adpositions

	Cat	Type	Formation*	Gender*	Numb*
1	Adposition				
2		preposit	simple fused	(m f)	sg pl

Only the Type Preposition is valid for Italian Adpositions. The Formation feature indicates whether a preposition is simple or fused with a definite article. This information is only relevant for those prepositions which allow fusion with articles (*a, di, da, in, con, su*). Gender and Number refer to the gender and number of the fused articles.

Conjunction

	Cat	Type	Coord-T	Subord-T
1	Conjunction			
2		coord subord		
3			simple initial no-initial correlative	+infve compar +fin

The Category **Conjunction** is obligatory.

Recommended information is:

- **Type** (coordinating and subordinating).

Optional information is:

- **Coord-T** provides distinctions among coordinating types, these are language specific (the proposed values are 'simple', for conjunctions between conjuncts, 'initial' for the first conjunction in repetitive constructions, 'no-initial' for following conjunctions and 'correlative' (introduced for Spanish),

- **Subord-T** provides distinctions among subordinating conjunctions: conjunctions which require a finite verb (+fin) a non-finite verbs (+infve) or which introduce a comparison (compar).

Information for Danish

	Cat	Type	Coord-T	Subord-T
1	Conjunction			
2		coord subord	simple initial no-initial	+infve compar +fin

Information on Type, Coord-T and Subord-T applies to Danish conjunctions.

Information for English Conjunctions

	Cat	Type	Coord-T	Subord-T
1	Conjunction			
2		coord subord	simple initial non-init	+infve compar +fin

Information on Type, Coord-T and Subord-T applies to English.

Information for Italian Conjunctions

	Cat	Type
1	Conjunction	
2		coord subord

In Italian, the same elements can often have the function of conjunctions, prepositions and adverbs.

There is no agreement concerning the inclusion of some elements in the conjunction or in the adverb class (or in them both). The Type distinction can be recommended, but subtypes of the coordinating and subordinating conjunction could also be introduced.

Numeral

	Cat	Type	Gend*	Numb(*)	Case*	Function
1	Numeral					
2		cardinal ordinal	(m/f/n/c)	sg pl		

Numerals are not always recognised as an independent category. In the EAGLES specifications it is possible to treat them as a category or as a type of Pronoun, Determiner or Adjective.

Recommended information is:

- **Type** (cardinal, ordinal).
- **Function** is introduced to account for systems which distinguish between Pronouns and Pronominal Adjectives (Italian) and those which distinguish Determiners and Adjectives (French, GENELEX).
- Gender and **Case**, are inflectional features relevant to some languages.
- **Number** may or may not be an inflectional feature depending on the language and on whether the items are cardinals or ordinals (ordinals do not have a value for number) thus we have put the star in parenthesis.

Information for Danish Numerals

	Cat	Type	Gend*	Numb(*)
1	Numeral			
2		cardinal ordinal	n/c	sg pl

In Danish, Numerals can be considered as a subclass of Adjectives or as an independent class.

A few numerals (*en, anden*) are inflected for Gender and Number.

All cardinal numbers except *en* (one) are, of course, plural in Number and this is not an inflectional feature.

Information for English Numerals

	Cat	Type	Numb
1	Numeral		
2		cardinal ordinal	sg pl

Only the Type and the Number attributes apply for English. The Number distinction, when applicable, is inherent in each numeral.

Information for Italian Numerals

	Cat	Type	Gend*	Numb*	Function
1	Numeral				
2		cardinal ordinal	m/f	sg pl	pronoun determiner

Traditionally, Italian Numerals are considered a subclass of Pronouns and Pronominal Adjectives.

Numerals can function as pronominal Adjectives with two possible functions: pronoun or determiner.

Ordinals inflect for Gender and Number.

Other Categories

A number of small categories which were difficult to classify were mentioned in the EAGLES specifications. They are not included here in this specification, but suppliers should be required to indicate any such extra categories which they use in their documentation.

4. Subcategorisation information

In this section, the similar general approach as for Morphosyntax was taken for subcategorisation features. Here, the starting point are the EAGLES guidelines (Sanfilippo *et al.* 1996). As with the morphosyntactic specifications, the EAGLES approach to standardising subcategorisation in lexica was bottom-up, comparing a number of syntactic theories (GB, LFG, HPSG, Categorial Grammar and Dependency Grammar). This comparison revealed that the following basic notions were taken into account in all the theories:

- Argument Structure
- Grammatical Relations
- Control and Raising
- Expletives
- Morphosyntactic features of subcategorised for elements

In addition, the practices in 7 practical NLP lexica and 6 annotation schemata for tagging corpora were surveyed and used as input to a consensual definition of the specifications for subcategorisation information to be included in lexica.

The tracing of the notions crucial in lexical entries to represent information that concurs to define and discriminate a syntactic structure, draws inspiration from (i) the experience gained within the PAROLE project, where the EAGLES guidelines have been concretely applied to a set of twelve lexicons and (ii) from the work performed within ISLE, where the syntactic basic notions have been investigated for the monolingual level, but also in view of the multilingual transfer.

A general presentation of the lexical notions for this level of description, will be provided by means of examples. They will be also described either as complex notions but also when needed in terms of their constitutive sub-elements.

In EAGLES the notion of subcategorisation[7] is interpreted as as being "concerned with the lexical specification of a predicate's local phrasal context" and "referring to typical collocations sanctioned by strong syntactic/semantic selection (head/complement relation), thus leaving out other collocation types such as head/modifier, head/specifier etc.)" (Sanfilippo *et al.* 1996, p.1). This means, for example, that co-occurence restrictions between determiners and their head nouns, which might be encoded on a given determiner's lexical entry, are not treated here. Synthetically subcategorization corresponds to a set of possible syntactic structures (the head and its syntactic arguments, with their phrasal realization) associated with an entry (typically a verb, but also a so-called predicative noun, an adjective or an adverb). The probability to appear in a corpus with a specific syntactic context can be also specified.

4.1 Subcategorization Frame

To sum up, information about subcategorization can be expressed by means of a list of sub-elements and in this sense can be considered as a *complex basic notion*. Sub-elements are:

1. A list of slots/positions representing the syntactic arguments (mandatory or optional) and their phrasal realization;
2. Categorial and morphosyntactic constraints concerning the lexical unit being described (the *Self* in EAGLES terminology);
3. Surface order information;
4. Frame probability.

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

Not only verbs have a subcategorization frame.

In the case of nouns, both deverbal and non-deverbal nouns may take arguments. In deverbal nouns the arguments may be inherited from the verb from which they derive, as in the example below:

The Romans destroyed the city.
The Romans' destruction of the city

For non-deverbal nouns we also allow for the possibility of an analysis in which, for example, the prepositional phrases in the following examples are considered to be arguments of the noun.

a book of verse
the journey to Paris

Of course, whether a specific lexicon includes such an analysis is dependent upon the syntactic theory or approach which has been adopted.

The possibility to express in an explicit way the information inherent to the subcategorization frame of a lexical entry is crucial for the weight it can have from a multilingual point of view.

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

un abito di velo (a voile dress) vs.	*bassa marea (low tide) vs.*
un velo di tristezza (a veil of sadness)	*una marea di gente (a stream of people)*

Adjectives present a subcategorization frame as well. In the Italian extension of PAROLE, the CLIPS lexicon, it has been chosen to give one-argument frame to the adjectives that do not bear any complement (Ruimy & Monachini 2002), cf. e.g. *veloce* (*fast*), *bello* (*beautiful*), where the only slot of the frame is filled by the nominal head modified (or predicated) by the adjective. Manifold reasons ground this choice: at the syntactic level, including the nominal head in the frame allows to better specify the head itself, giving also information on the noun (e.g *contiguous* that modifies obligatorily a plural noun, e.g. *angles*); from the perspective of the linking between syntactic and semantic frames, it facilitates the correspondences between the predicate argument and the slot of the syntactic construction; finally, semantically, the predicate argument is immediately projected on to the syntactic slot headed by the deadjectival noun (e.g. *intelligent girl*; *girl's intelligence*). Some predicative adjectives present a syntactic frame with more than one slot. The second slot can be filled by a nominal complement *un ragazzo abile al lavoro* or by a clausal complement *una persona felice di fare qualcosa*[8]

[8] In such cases the coreference between the subject of the inifinitive and the adjective nominal head is marked.

The notion of subcategorization is strongly interconnected to the notion of argument structure (see below the section on semantics): they both lie at the heart of the correspondence between syntax and semantics. They have a strong discriminating power in sense disambiguation - and consequently in translation selection -, giving rise to different translation equivalents on the basis of the different thematic roles and semantic characterization a syntactic position can take.

The notion will be presented here only at the level of syntax, focussing on how the subcategorization is crucial to discriminate between syntactic structures of a same entry. The correspondence between syntax and semantics will be dealt with later, after the introduction of the basic notions for semantics.

The basic feature indicating that a lexical item subcategorises for certain elements is Frame. On the basis of the investigation and comparison of various existing practices, EAGLES has proposed a model of the frame. In order to clarify the feature checklists provided, the overall subcategorisation frame model is reproduced as Figure 1. below. The obligatory information (according to EAGLES) is indicated by underlining.

Frame[9]

id	*unique_key*				
Slot*	id	*unique_key*			
	index	*number*			
	optionality	*yes/no*			
	Realisation:				
		id	*unique_key*		
		controlled_by	*number*		
		obviates	*number*		
		in_order	*pre/post*		
		syntax:	category:		
				label: *np | vp | pp ...*	
				features *[feature_i value_i]**	
			function:		
				name *name_i*	
				subject *yes/no*	
				predicative *yes/no*	
		semantics[10]:	themat_rol:		
				rol_name *name_i*	
				rol_features *[(feat_i val_i)]**	
			semant_class:		
				class_name *name_j*	
				is_a *class_name_l*	
self	[features [*featurei valuei*] *]				
control	*control-verb|subject-raising-verb| object-raising verb*				

[9] The features are hierarchically defined so that the values of attributes may either be simple (leaf) values or more complex structured features. Thus the checklist of attributes is presented in a number of different linked tables. Simple leaf values are indicated by italics whilst values which are themselves structured are given in ordinary type.

[10] With respect to the frame presented here which is direclty taken from the EAGLES recommendations, this set of information is not dealt with here. Being the core of the syntax-semantics linking mechanism it will be devoted a separate section in the semantic part.

rel_order:		
	before_slot	*number*
	after_slot	*number*
	after_realisation	*id*
	before_realisation	*id*

Figure 1. The EAGLES Subcategorisation Frame Model

Each element of the model is devoted a special section below.

4.1.1 Slots

Slots are the subcategorized elements of the syntactic frame (the *syntactic positions* in the GENELEX/PAROLE terminology) are specified as to information described below.

Slot				
	id	Index	Optionality	Realisation
obligatory		*number*	*yes/no*	Syntax
optional				Semantics[11]

- **id** is a unique identifier

4.1.2 Index

Index is a number indicating the canonical **ordering** of the slot.

The slots of the subcategorization frame have a conventional or canonical order that can be different from the linear order of the positions in real sentences, since the surface order is not something that should be encoded in the lexicon. Anyway, as stated in the recommendations on Subcategorization (Sanfilippo *et al.* 1996), *"for some lexical units and for some languages…some verbs may constrain the possible order of their slots or slots realizations more than others*".

The information about linear order can be important: for example, in Spanish and in Italian, the position of the adjective as pronominal or postnominal (or both) encoded in the lexicon has consequences on the sense distinction, (i.e. *pobre hombre/pover uomo – unhappy, miserable man* – is different from *hombre pobre/uomo povero – poor, lacking money man* –).

In the document of Subcategorization the possibility to deal with lexically specified constraints on order (over and above the constraints imposed by the grammar) is provided. It has a set of complex values as shown in the following table and is imposed with a progressive number (starting from 0).

Rel_order:	
before_slot	*number*
after_slot	*number*
after_realisation	*id*
before_realisation	*id*

[11] Semantic Realization of Position is dealt with in the section devoted to the Linking between syntax and semantics.

4.1.3 Optionality

In many cases, there is the need to state the optional realization of a syntactic slot within a subcategorization frame. In order to assess the optionality e.g. of a verb argument, 'nuclear' sentences should be considered , in a 'not-marked' context (since marked context can admit even the omission of traditionally obligatory complements). For the verb *to sing*, the structure *you are singing* can be considered self-explanatory, whereas, for the verb *to buy*, *you are buying* is retained as needing an obligatory direct object for the completion of the sentence[12]. Optionality, in a monolingual framework, can turn out to be a clue for sense disambiguation, e.g. a literal meaning vs. a figurative reading: *la legna si accese* (*incendiarsi*) vs. *Gianni si accese d'ira* (*adirarsi*)[13]. The same can be true for nouns, e.g., *I lost my key* (Instrument) vs. *to know the key* (Solution) *to the enigma*, where the abstract sense obligatorily requires the presence of the slot pp-*to*.

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

- the absence of a slot excludes the presence of another slot : cf.

  John refuses obedience to Mary/John refuses obedience/John refuses
  but not **John refuses to Mary*

where the absence of the direct object prohibits the presence of the indirect object.

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

  John competes with Mary for the exam/John competes for the exam/John competes with Mary
  but not **John competes*

where the presence of one of the two slots is needed in order for the sentence to be acceptable.

4.1.4 Slot Realization

This is the place where the phrasal realization of the syntactic argument can be specified (saying for example that the first slot, Slot*0* – or in PAROLE terminology, Position*0* – is instantiated by a Noun-Phrase. etc.).

	Category		Function	
obligatory	Non terminal	*S* (sentence)	Name	*name*
		VP (verb phrase without subj)		
		NP (nominal phrase)	Subject	*yes*
		PP (prepositional phrase)		*no*
		AP (adjectival phrase)		
		ADVP (adverbial phrase)	Predicative	*yes*
		XP (under-specified phrase)		*no*
Optional	Terminal	Morphosyntactic features		

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

[13] Additionally, in a multilingual perspective, this can imply different translations: *the wood caught fire vs. John blew up with rage*.

The syntactic properties of a slot realization can be expressed by means of *terminal* or *non-terminal categories*.

4.1.4.1 Non-terminal categories

The list of **non-terminal categories** in the figure above is the same as proposed in the EAGLES Recommendations (Sanfilippo *et al.* 1996, pp. 64-65).

The EAGLES phrasal category labels are provided as a generalised set of features from which the lexicon developer can choose. The labels are intended to be general and can have sub-types. So, for example, different types of clause such as those introduced by a complementiser are assumed to be subsumed under S (sentence). In addition, EAGLES also suggested the category DETP, however given that specifier/head information relations are not considered as part of the subcategorisation frame, we have left this possibility out.

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

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

It is also possible that certain predicates subcategorise for specific parts of speech rather than phrases or clauses. For such terminal categories the same labels as those used for morphosyntactic distinctions should be used.

4.1.4.2 Terminal categories

The list of **terminal categories** (the object SyntagmaT of PAROLE), are those provided by the EAGLES Morphosyntax Group (Monachini & Calzolari 1996):

N- Noun
A- Adjective
P- Pronoun
V- Verb
ADV- Adverb
CNJ- Conjunction
ADP- Adposition
DET- Determine
ART- Article
INTJ- Interjection

Besides grammatical category and functions, slots can also be characterized using **restricting features**, i.e. labels that allow to specify further restrictions of morphological kind (i.e. tense, mood, gender, etc…) or lexical kind (for example the lexical introducer of a prepositional phrase).
Since the same features can be used to characterize the information about the head of the construction (the *Self* in the EAGLES terminology) as well, they will be dealt with in the section Restricting features.

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

4.1.4.3 Function

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

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

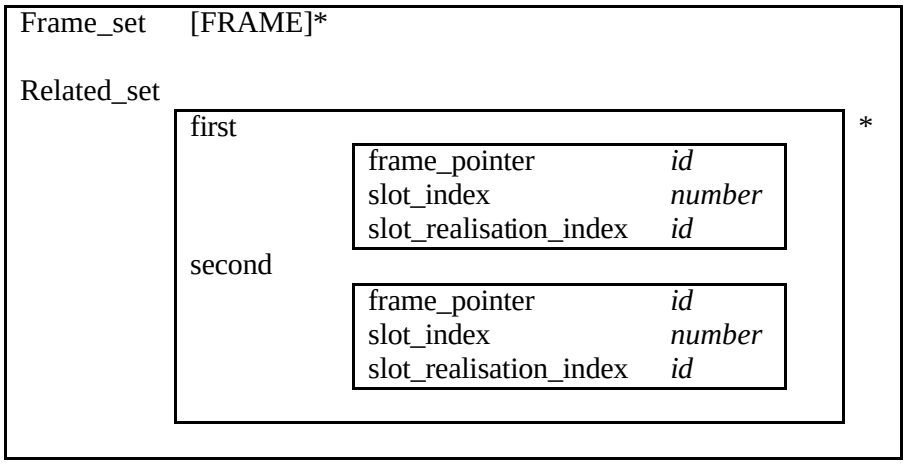
The grammatical function characterizing one of the syntactic positions of the frame turns out to be a crucial notion. At multilingual level, for example, it can be constrained adding information and expressing, for example, a typical object or subject of a verb. *Typical subject* or *typical object* very frequently act as sense indicators. As an example the Italian verb *dondolare* gets different meanings and translations according to the different typical objects: *to swing one's arms*, *to dangle one's feet, to rock the cradle*.

4.2 Regular Syntactic Alternations and Frameset

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

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

The figure below shows how the device works.



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

A Frame_set is a set of related frames in which different surface alternations can be explicitly linked within related sets of frames, where the correspondences between different slots in related frames are indicated via use of the same value.

Each lexical item has an associated **Frame_set** which may include only one frame (**frame_n**)

Frame_set:

	id	name	frame_n	related_n
Obligatory	*id*		*id_ref*	
Optional				Related

The obligatory information is:

- **id** - the unique identifier of the frame set
- **name** the name of the frame set
- **frame_n** (standing for all the possible frames: frame_1, frame_2 etc.) the identifier of the frame in question

A frame set can contain only one frame (and thus there is no alternation, therefore the following attribute is optional and only applies to cases of alternation:

- **related_n** (standing for possibly more than one related pair: e.g. related_1, related_2 etc.). The value of related_n is itself structured, as shown in the next table.

Related:

	id	name	first_slot_pointer	second_slot_pointer
Obligatory	*id*	*name*	Slot_pointer	Slot_pointer
Optional				

All the features of Related are obligatory:

- **id** the unique identifier of the object Related.
- **name** the name of the object Related.
- **first_slot_pointer, second_slot_pointer** these two attributes indicate the two slots which are related. They both have the same type value (Slot_pointer), which is structured as shown in the next table.

Slot_pointer:

	frame_pointer	slot_index	slot_real_index
Obligatory	*id*	*number*	*id*
Optional			

All the features are obligatory:

- **frame_pointer** the id of the frame (out of the frames assigned to the lexical item in its Frame_set).
- **slot_index** the number of the slot in the frame pointed to.
- **slot_real_index** the id of the realisation of the slot pointed to.

4.3 Frame Probability

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

4.4 Self

The Self encodes the peculiar features and restrictions of the lexical entry in the specific syntactic context it appears. For verbs, a particular behaviour with respect to the application of grammatical rules, e.g. a transitive verb not passivizable, subclass of a verb, auxiliary selection, passive voice inibition, etc.; for nouns, countabilty, morphological restrictions such as use of gender or number; for adjectives: attributive/predicative function, adjective position with respect to the nominal head, gradability; for adverbs: semantic subclass and modified part-of-speech. This information is very useful at the multilingual level, when it addresses the translation in a specific direction.

Very important is the possibility to specify complex heads in order to represent polylexical units. A complex head is something having an inner structure made of embedded positions describing the multiword components. This necessity strongly arises during the phase of entries creation, when it is important to have at disposal a device to represent in a straightforward way an entry like "make an impression" (complex head formed by *make* -verbal head- + a slot for the NP "impression").,

4.5 Restricting Features

The information about the syntactic frame and the syntactic behavior of an entry can be further specified by means of a set of features. In most cases, the only use of categories is not sufficient to supply the necessary information and, categories must be completed by using restricting features.
The EAGLES Documents on Subcategorization (Sanfilippo *et al.*, 1996) and on Morphosyntax (Monachini & Calzolari, 1996) provide a classification of the possible types of information that can be used to refine the information already specified in the Slots and in the Self.
Features are distinguished in (i) morphosyntactic and (ii) lexical.

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

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

  collezionare francobolli (to collect stamps)
  pullulare di stelle (to swarm with stars)

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

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

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

Lexical features, on their turn, help to describe various aspects of the lexicalization of a phrase (its preposition etc.) and are also crucial at multilingual level, since we may need to select a specific preposition within a subcategorization frame.

4.6 Control

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

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

In raising constructions (cf. Sanfilippo 1996, p.81), the subject expressed in the governed sentence is "raised" as subject of the governing verb[16].

sembra che Luca sappia l'inglese (It seems Luca can speak English) ➔
Luca sembra sapere l'inglese (?Luca seems to be able to speak English).

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

The feature controlled_by is only relevant to frames in which control occurs and refers to the slot which has the control.
Unlike the EAGLES recommendations, this feature has been made optional rather than obligatory. Whilst in cases of control we strongly recommend that it is included, there are some cases where it could be impossible to determine the control relation without explicit reference to the entire context in which the predicate occurs.

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

5. Semantic Information

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

The previous EAGLES guidelines in the area of lexical semantics have been hence re-interpreted under this perspective, trying to provide the set of information necessary to be dealt with at this level of representation. In this light, the bulk of semantic information encoded in the SIMPLE lexicons (that, built on the EAGLES recommendations, has been taken as the basis for the analysis carried out here) are also re-examined and integrated (with other dimensions coming e.g. from WordNet). Other realities have been taken into account, since the notion of *word meaning*, which is central to semantics description, is not uncontroversial. In the lexicographic tradition, the word meaning is the *sense*, the unit resulting from the subdivision of the lemma in its readings. In lexicons *à la* GENELEX (or SIMPLE), the word meaning is represented by the SemU – the Semantic Unit – corresponding to the traditional notion of word sense and constituting the nuclear building block of the whole semantic description. It is the semantic unit that is linked to a given ontological type, it is the semantic unit that the semantic frame is associated to, and it is the semantic unit that, alternatively, works as the target and the source of all semantic relations. A different modality of representation resorts to the *synset*, the *set of synonyms* that constitutes the building block in WordNet (Fellbaum, 1998) and WordNet-like kind of resources (Vossen, 1999). During the years, WordNet has become an outstanding reality for the lexicon community, with WordNets dedicated to dozens of languages and used in a wide variety of applications. Thus, it is important to take WordNet and its basic structure into consideration, ensuring that all the already encoded resources could be easily mapped into the standard being designed.

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

5.1 Semantic Frame

This is a complex notion, that specifies the predicative argument structure of a lexical unit described in terms of the following types of sub-elements: the predicate, which on its turn is described by means of a list of arguments, their semantic role and the selectional restrictions the predicate operates on them. This notion "incorporates most of the lexical semantics elements, since predicates are often the 'kernel' of propositions" (Sanfilippo *et al.* 1999). In SIMPLE, the semantic frame is recommended and instantiated with a very high degree of detail (Lenci et al. 2000b, p. 46).

In a multilingual perspective, it is the place where many operations necessary to go from one language to another occur: all information connected to the semantic frame helps such operations. Information about the type of link between the predicate and the semantic unit can have repercussions on cross-language linking as well.

5.1.1 Predicate

The information about the predicate is relevant for verbs, predicative nouns, adjectives, prepositions and adverbs.

The approach to predicate can be of two types: multilingual, as language-independent primitive predicates, or monolingual, as language-dependent lexicalized predicate. On the one hand, 'abstract' predicates to be shared by homogeneous classes of semantic units across languages could acquire a kind of "interlingua" valency (the abstract predicate PredPROPERTY_OF could be linked to all Property denoting nouns, such as *bellezza, beauty, beauté; altezza, height, hauteur,* … independently of lexicalization in every language).
EAGLES recommends (and SIMPLE instantiates) language-dependent lexicalized predicates which present "the advantage of reducing the complexity of the linking with syntax" (Lenci et al. 2000b, p.46).

Predicative entries are ascribed a *semantic predicate*, being provided with the so-called predicative representation. In SIMPLE, the approach adopted for the selection of predicates foresees that members of a whole derivational paradigm are all linked to the same predicate. It follows that different semantic units may share the same predicate in the predicative representation: e.g. the verb *destroy* and the nouns *destruction* and *destroyer* all point to the PredDESTROY; similarly, the verb *employ*, and the nouns *employment*, *employer* and *employee* are linked to the PredEMPLOY; the deadjectival noun *intelligence* and the adjective *intelligent* share the PredINTELLIGENT.

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

- Verbal lexical units, such as *employ* and *destroy* present with respect to their predicate (PredEMPLOY PredDESTROY) a MASTER type-of-link, which stands for 'the priviledged lexicalization of the predicate';
- *employment* and *destruction*, on their turn, constitute EVENT NOMINALIZATION (whose surface realizations instantiate all the arguments of the relevant predicate)[17].
- *Employer* and *employee* are, respectively, AGENT and PATIENT NOMINALIZATION of PredEMPLOY. Within the type of link there is also the possibility to specify that in both nominalizations the phenomenon of 'argument absorbtion' takes place, i.e. *employer* absorbs in the lexical head the ARG0:agent, whereas *employee* encapsulates ARG1:patient.
- INSTRUMENT NOMINALIZATION and locatives (OTHER NOMINALIZATION) are ascribed the relevant predicate as well, cf. *mixer* that incorporates ARG2:instrument of the PredMIX and *breeding* that realizes ARG2:location of the PredBREED.

5.1.2 Arguments

The notion of predicate involves the specification of the number and type of arguments. Arguments as well as predicates are 'lexically driven', so each predicate has its 'own' arguments. Determining the list of arguments involved in a predicate is not a trivial task. As an example, SIMPLE states that the choice of the number of arguments for a predicate has to be determined on purely semantic grounds: it is perfectly possible for a semantic argument not to be mappable to any syntactic position, and, conversely, it is perfectly possible for a syntactic position to remain unlinked to any argument.
At multilingual level, arguments represent a critical notion, since most of the transfer operations seem, principally, to affect aspects of the syntactic facet connected to a semantic frame, the number of arguments involved in Frame1 and Frame2, the order of the slots filled at the level of surface syntactic realization.

[17] The fact that the verbal and deverbal noun structures share the same predicative representation can be of extreme utility in order for, e.g., the two different surface realizations linked to the PredDESTROY (*la distruzione della città da parte dei nemici* --the destruction of the city by the enemies - and *i nemici distruggevano la città* - enemies destroyed the city) be recovered.

5.1.3 Thematic Roles

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

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

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

- Agent
- Patient
- Experiencer
- Location
- Instrument

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

Pred*give*: ARG0-Agent ARG1-Patient ARG2-Beneficiary

Gianni dà un libro a Maria (pp-a) *John gives Mary* (np) *a book*

5.1.4 Selectional Restrictions

Selectional restrictions should rather be intended as *selectional preferences* (Sanfilippo *et al.* 1999, Lenci et al. 2000b and Calzolari *et al.* 2001), as arguments which are *preferably* selected by a predicate.

Selectional restrictions on arguments can be specified in terms of the following types of information:

- *Semantic Type*, taken from the list of semantic types that form the Ontology (cf. Semantic Type);
- *Features* or *Notions*, e.g. a set of semantic types (Human Animal, i.e. the ∪ of the set of Humans and the set of Animals), a semantic type plus feature(s) (Human +FEMALE) .
- *Semantic Unit*: for instance, *bark* has a two-argument semantic frame, where the second is restricted to *dog* (where *dog* should include all instances of class DOG).
- *Synsets*: restrictions can be enforced also by means of a group of admitted synonyms[18].
- *Collocations:* restrictions can involve a lemma typically accompanying the unit at hand.

Restricting the predicate's argument by means of semantic features allows to overcome cases in which the use of other expressive means, e.g. semantic types, seem to fail in capturing the full range of arguments, being, alternatively, too wide or too restrictive[19]. Features, which cut across the type hierarchy, allow in fact to capture a more suited set of lexical units and are considered more powerful in identifying preferences: cf.

[18] Even if it should be taken into account that not always members of a same set of synonyms can be perfectly interchangeable.
[19] Selecting the type Human for the agent of the Pred*eat* excludes Animal, whereas Living_Entity covers also undesiderable Vegetal_Entity.

the restriction on patient of the Pred*eat*, that excludes vegetals and fruit if expressed with the type Food, whereas captures also other semantic units distributed over different semantic types (Vegetal, Fruit, Vegetal_entity, Substance, Natural_Substance …) if expressed by the feature [+edible] (cf. distinctive features).

5.1.5 Synset

The *synset* is the set of synonyms that plays the central role of *lexical concept* in WordNet. Following psycholinguistic assumptions, the idea is that the human lexical memory is organized around concepts that words can be used to express. The same meaning can thus be carried by more than one word and represented by the group of those words themselves.
This is an important shift from the lexical organization discussed above: the synset can be viewed as a set of *senses* of different lemmas (the *variants*, in the EuroWordNet terminology, the SemUs in GENELEX-SIMPLE terminology) grouped on the basis of their reciprocal synonymy. The following list of word senses are examples of two actual WordNet1.6 synsets obtained with the search word *home*:

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

The synset is the node of the semantic net, that works as an anchor for every semantic relation.

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

- The synset with one or more synonyms (variants, senses, SemUs) as sub-elements and characterized by the following attributes:
  - POS indicator (*mandatory*)
  - Gloss (*optional*)
  - Example (*optional*)
- A list of one or more relations. The relations can be of different types, representable by means of different attributes: monolingual semantic relations, equivalence crosslingual relations and plug-in relations[20] between generic and domain-specific wordnets.
- Features providing the semantic and ontological types.

5.1.6 Features

Semantic Type

Semantic type appears to be a crucial notion, since it establishes a link between a word-sense and an ontological type system which is used to classify senses themselves, thus allowing to assign it to a specific position in the nodes of the type hierarchy: *dog* [Animal ← LivingEntity ← ConcreteEntity ←…]. In cases where senses are not defined on the basis of an ontology, the semantic type can be also obtained via semantic hyperonymic relations with another word-sense, *dog* isa *animal.*
This notion is uncontroversial (even if there is no agreement on a unique system of semantic type/ontology): the semantic type of a word sense is a means to discriminate among other possible senses of the same lemma. Looking at well-established practices in computational lexicons or Machine Readable Dictionaries, all of them make use of it (Calzolari *et al.* 2001). This notion is considered as *required* by SIMPLE (Lenci et

[20] As instantiated in the ItalWordNet databases (Roventini *et al.* 2002).

al. 2000b, p.37), i.e. it is part of the core information included in the minimal requirements for computational lexicons at semantic level[21].

Domain

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

Distinctive Features

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

Semantic Relations

Together with the above expressive devices, the semantic purport of an entry is also represented by means of semantic relations between two semantic units[22] (senses). Relations can also be established between synsets, as in the WordNet model[23].

Information that traditionallly is committed to relations consists in meronymy – *part_of* (finger, hand) –, and its inverse relation holonymy – *has_part* (carburettor, car) –, antonymy, with its variuos types of opposite relations – (true, false); (hot, cold) – as discussed in Cruse, 1986. The utility of such dimensions in various types of applications is carefully reported in the EAGLES Recommendations on Lexical semantics (cf. Sanfilippo *et al.* 1999, p. 238).

In the framework of the SIMPLE experience, relations between SemUs are used to instantiate traditional Qualia roles of the Generative Lexicon (Pustejovsky, 1995). This allowed lexicographers to represent the richness of semantic relations in natural language and, at the same time, to capture the essence of a word meaning. In addition, the set of Qualia roles has been made richer and simultaneously stricter. Richer because each of the four Qualia roles has been represented in the form a relation, which is in turn the top of a hierarchy of other more specific relations. Stricter in that the enlarged set of relations allow to capture more fine-grained relations holding between different senses. These hierarchies of relations (specifically 64 semantic relations, cf. Appendix C) within the four Qualia have been called *Extended Qualia Structure*, (cf.

[21] The SIMPLE and Top EuroWordNet Ontologies are included here as examples of commonly agreed-on semantic type systems (cf. Appendix A and B).

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

[23] In this case, we speak about *lexical* relation.

Lenci et al. 2000b, pp. 59-71). Qualia relations, combined together, characterize, indeed, semantic types of different degrees of complexity and concur to maintain the (Qualia) structure of a semantic type. Relations have been also given a weight, depending on their being type-defining with respect to a semantic type or not.

Derivational relations (*beauty*; *beautiful*) and regular polysemous classes (Animal/Food: $lamb_1$, $lamb_2$; Substance/Color: $turquoise_1$, $turquoise_2$) have been implemented as relations between semantic units as well.

In EuroWordNet the device of relations is used to represent relations holding between different set of synonyms (cf. Appendix D).

Collocations

Collocations, which EAGLES defines a kind of "word co-occurrence relations" (cf. Sanfilippo *et al.* 1999, p. 240), are crucial to define the semantic purport of a lexical entry which selects a particular meaning when it co-occurs with a given word. In collocations, the way words go together seems idiosyncratic and unpredictable: the selection operates at the lexical level rather than at general semantic level. This has a particular impact in multilingual operations in order to arrive at the correct translation equivalence in another language. Collocations can, by their nature, be encoded by means of the expressive device of relations, where the typical collocate of a word is the target of the relation[24]. EAGLES provides a set of information generally necessary to be specified for collocations (cf. Sanfilippo *et al.* 1999, p. 245): direction, word-distance, dependency, dependency type, probability.

5.2 Linking Syntax and Semantics

The type of notion dealt with in this section refers to one of the most crucial aspects of computational lexicons, which goes by the name of *linkage of syntactic and semantic levels*.

This operation consists in relating the *semantic frame* pointed by a semantic unit and the *syntactic frame* the latter is associated with, specifying how *semantic arguments* and *syntactic slots* correspond each other, i.e. how arguments are instantiated in the surface.

In SIMPLE a battery of rules to map the semantic predicate onto its possible syntactic surface instantiation(s) has been defined.
Rules are able to deal with typical cases of:

- isomorphism, where slots and arguments correspond to each other in number and range (mono- bi-, tri-, tetra- valent ISOMORPHIC correspondences: ARG0-SLOT0; ARG1-SLOT1 …),
- correspondence between slots and arguments appearing in crossed order (CROSSED correspondence: cf. *destroy* and *destruction*: ARG1-SLOT0; ARG0-SLOT1),
- non correspondence between syntactic slots and predicate arguments:
  - the case e.g. of adjuncts which are part of the syntactic frame but extraneous to the semantic one (REDUCED correspondence) or, conversely,
  - semantic arguments that do not appear in surface realizations (e.g. 'Meteorological' predicates [*snow*] *snowed*) or can be lexically encapsulated[25] (AUGMENTED correspondence).

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

[25] If considered in multilingual perspective, argument encapsulation has interesting implications, when dealing with cases of predicates which behave differently, across languages wrt this phenomenon, cf. Eng. *to funnel* – It. *versare con l'imbuto* and Eng. *to hammer* – Fr. *enfoncer avec un marteau*.

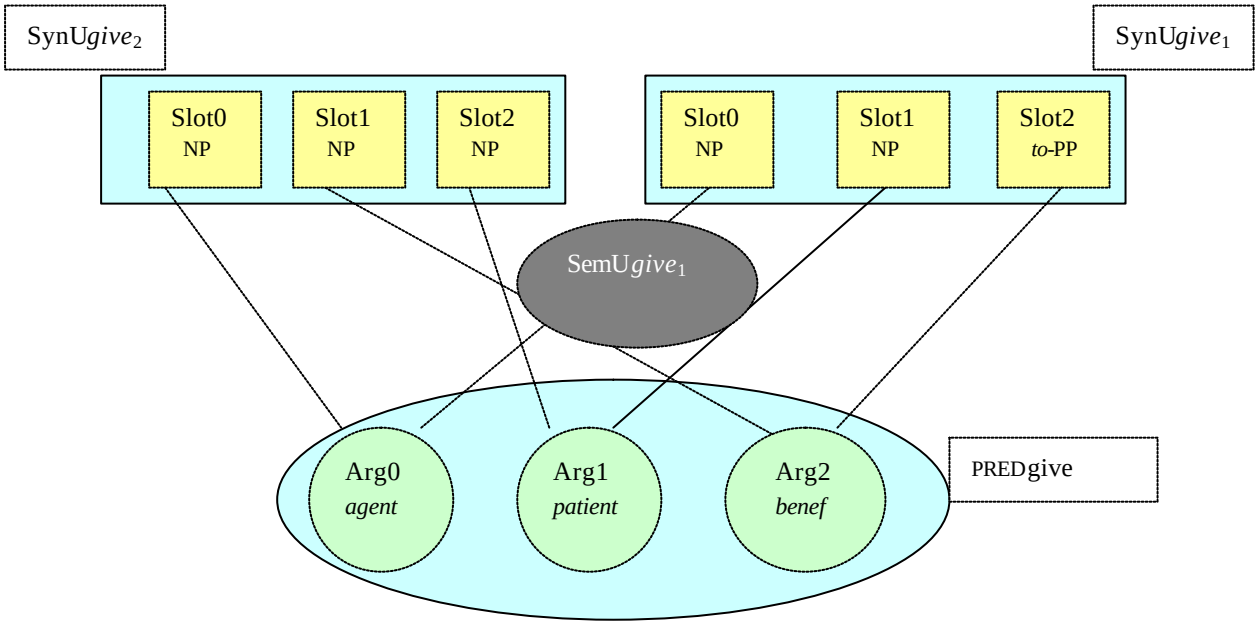
To give but an example of the usefulness of the mapping rules and just a flavour of how they work, a case of regular dative alternations is taken into consideration:

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

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

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

Both are associated to the same semantic unit <give>[ChangePossession] which points to the predicate PredGIVE(Arg0:agent, Arg1:patient, Arg2:beneficiary). Starting from this predicative representation, the two alternating surface realizations can be reconstructed by way of the appropriate mapping rules.



crossed correspondence ----------
isomorphic correspondence ----------

Different mapping rules will account for the differences in correspondence between the predicative structure and the two possible surface instantiations: the arguments of PredGIVE, on the one hand, are associated with the slots of the syntactic unit *give*$_1$ through an *isomorphic* correspondence (Arg0→Slot0, Arg1→Slot1 and Arg2→Slot2) on the other hand, will be mapped on to *give*$_2$ syntactic frame via a *crossed* correspondence (Arg0→Slot0, Arg1→ Slot2, Arg2→Slot1).

6. Multilingual Operations

The presentation of the basic notions for the multilingual part takes inspiration directly form the experience gained within ISLE (Calzolari *et al*. 2002) where a common model to represent multilingual content within resources is identified. ISLE provides such a common model analysing various approaches, i.e. i) direct architecture consisting of simple *word-to-word* replacement, ii) transfer approach exploiting the syntactic and semantic representation of the source and the target languages to go from L1 to L2, or iii) interlingual approach**,** based on the idea that translations from SL to TL should pass through a language independent representation.

If at the monolingual level basic notions mostly concern "static" lexical objects (such as syntactic slots, semantic arguments, restricting features etc.)[26], from a multilingual perspective basic notions involve the set of operations that use these very lexical objects as *arguments*. In the MILE, the information about the syntactic and semantic behavior of an entry is constrained (adding or deleting semantic and/or syntactic information) by means of a set of transfer conditions that allow to create correspondences between language pairs. In other words, all information concurring to define a syntactic structure or a word meaning from a monolingual point of view can be exploited for multilingual requirements and, together with the transfer conditions, can be regarded as *basic notions.*

As far as the multilingual layer is concerned, among the most important above-mentioned reference works for ISLE we find i) the "Rapport sur le MULTILINGUISME" of the GENELEX Consortium (1994) and ii) the transfer operations of OLIF (Thurmair, 2000), the interchange format used in many industrial MT systems. ISLE extends the GENELEX model towards the definition a more flexible framework where different approaches can be instantiated, in particular opening the door to an interlingual approach. With respect to the objects presented in the GENELEX multilingual layer, "new" basic notions have been introduced coming from the monolingual layers, to be exploited at the multilingual level as well, i.e. the *synset* – that can be used in cross-language correspondences – and the semantic relations – on which the transfer mechanism operates in the same way as on other notions. Even if ISLE takes inspiration mostly from a transfer-based multilingual model, in the model proposed it should be possible to represent and instantiate, in addition, also a more elementary and a more conceptual/abstract multilingual model:

- the direct transfer architecture can be instantiated recurring to the simplest and immediate correspondence, i.e. that between morphological units;
- the interlingual approach to translation can be implemented, exploiting and specializing the semantic/conceptual level: the monolingual notion of lexical predicate can be extended to a more abstract notion of non-lexicalized predicate, where abstract primitives can be combined to realize a language independent, neutral and conceptual representation. In this sense, the representation resides outside the monolingual descriptions and does not need transfer rules, since the same internal representation is used for both the source and the target languages.

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

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

[26] cf. ISLE Deliverable, D3.1.

- *multilingual correpondences* that intervene in the linking process of monolingual lexical objects. Correspondences should be possible between:

  - morphological units pairs
  - syntactic unit pairs: this correspondece allows to put into relation two syntactic units independently of their semantic realization. Sub-element of this kind of correspondence is the correspondence between each slot of the SL and TL syntactic frames.
  - slot pairs: this correspondence allows to link slots of the descriptions attached to each syntactic unit. It should be possible to constrain or prohibit the realization of a slot, to force it to a given syntagma. The syntagma, on its turn, should be constrained and new slots added to the already existing list of slots and again constrained.
  - semantic unit pairs: when a correspondence is established between SL and TL semantic units, all the syntactic units connected to them are related, and implicitly, via the correspondence between syntax and semantics, their syntactic frames are linked as well. When predicative semantic units are put into correspondence, obviously their respective semantic frames are related as well.
  - predicate pairs: this correspondence allows to associate the predicates of each language, independently of the semantic unit(s) they are pointed by and, hence, independently of the semantic frames they are linked to.
  - argument pairs: it specifies the correspondences between arguments of the semantic frames of the SL and TL. It should be possible to add a semantic feature in order to better specify the argument or operate a constraint in order to cover the semantic gap, if any, between two elements in correspondence. It should be possible also to specify optional arguments which do not present any correspondence in the other language, or, conversely, to add arguments.
  - mixed pairs of semantic and syntactic units: allows to exactly specify which syntactic descriptions are linked for a given lexical meaning.
  - synsets: the notion of synset is not the most suitable in a MT system, since each member of the synset can have a different syntactic and/or collocational behaviour in generation with respect to other members. Moreover, it is not possible to realize a cross-language *variant-to-variant* mapping by using the synset (this correspondence is feasible only between word senses). The multilingual extension of a monolingual wordnet-*like* lexicon is, however, important for a range of cross-languages applications, such as CLIR, CLIE and CRQA.

- operations that can be used in the test and action mechanism.The core of the transfer is the mechanism of tests and actions of "*if…then*" type which apply respectively to source and target lexical objects. Operations can be of two types:

  a. *"Constrain" operations*: they apply to source lexical objects (test operations) and to target lexical objects (action operations). By means of this family of operations it is possible to perform a restriction on the value of syntactic and semantic elements, forcing for example a slot of the syntactic frame to be realized by a certain phrase. Subtypes of constrain operations are Constrain (Self), Constrain (Slot), Constrain (Syntagma), Constrain (Argument).

  b. *"Add" Operations*: they operate simply by adding the information individuated in the translation process to arrive to the correct equivalent. Subtypes: Add (slot), Add (argument), Add (syntagma), Add (Syntactic Feature), Add (Semantic Feature), Add (Semantic Relation)

References

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

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M. (eds.) 2002. Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entries), ISLE Deliverable 2.2 &2.3 CLWG, Pisa. http://lingue.ilc.cnr.it/EAGLES96/isle/.

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

Erjavec, T. & Monachini, M., (eds.) 1997. Common Specifications and Notation for Lexicon Encoding of Eastern Languages. Deliverable 1.1. Multext-East Project, COP-106

Fellbaum, C. (ed.) 1998. WordNet. An Electronic Lexical Database, Cambridge, Cambridge, The MIT Press.

Fillmore, C. J., Wooters, C., and Baker, C. F. 2001. Building a Large Lexical Databank Which Provides Deep Semantics. In Proceedings of the Pacific Asian Conference on Language, Information and Computation, Hong Kong.

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

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.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., and Zampolli, A. 2000a. SIMPLE: A General Framework for the Development of Multilingual Lexicons. International Journal of Lexicography, 13 (4): 249-263.

Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., Zampolli, A., Guimier, E., Recourcé, G., Humphreys, L., Von Rekovsky, U., Ogonowski, A., McCauley, C., Peters, W., Peters, Y., Gaizauskas, R., and Villegas, M. 2000b. SIMPLE Work Package 2 – Final Linguistic Specifications, Deliverable D2.2, workpackage 2, LE-SIMPLE (LE4-8346).

Monachini M. 1995, *Common Specifications and Notation for Lexicon Encoding of Eastern Languages*, COP Project 106 MULTEXT-EAST, WP-1, Task 1.1, Del 1.1., Pisa, 1995.

Monachini M., Calzolari N., 1996 *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages.* EAGLES Document EAG-LSG/IR-T4.6/CSG-T3.2, Pisa, Italy.

Pustejovsky, J. 1995. The Generative Lexicon, Cambridge, The MIT Press.

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

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

Ruimy N. & Monachini, M., 2002. Specifiche Linguistiche e Manuale di codifica – Livello Sintattico, Rapporto Tecnico del Progetto CLIPS del MURST, Pisa

Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M., Calzolari, N., Zampolli, A. 2002. CLIPS, a Multi-level Italian Computational Lexicon: a Glimpse to Data. In Proceeding of the LREC2002, Las Palmas de Gran Canaria, Spain: 792-799.

Ruimy N., Monachini, M., Gola, E., Calzolari, N., Ulivieri, M. Del Fiorentino, M.C., Ulivieri, M., Rossi, S. forthcoming. A Computational Semantic Lexicon of Italian: SIMPLE. In Linguistica Computazionale, Pisa, Giardini Editori.

Sanfilippo *et al*. 1996 *Subcategorization Standards. Report of the EAGLES Lexicon/Syntax Group.* SHARP Laboratories of Europe, Oxford Science Park, Oxford, UK.

Sanfilippo *et al*. 1999 *Preliminary Recommendations on Lexical Semantics Encoding.* Final Report, SHARP Laboratories of Europe, Oxford Science Park, Oxford, UK.

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

Underwood N. L., & Navarretta C. 1997. *A Draft Manual for the Validation of Lexica, Final Report.* Report submitted to ELRA under the validation task contract.

Vossen, P. 1999. Introduction to EuroWordNet. Computers and the Humanities, 32: 73-89.

Appendix A – The SIMPLE Ontology

General Ontology for Nouns and Verbs

1. **TELIC** **[Top]**
2. **AGENTIVE** **[Top]**
   - 2.1. **CAUSE** **[Agentive]**
3. **CONSTITUTIVE** **[Top]**
   - 3.1 **PART** **[Constitutive]**
     - 3.1.1. BODY_PART **[Part]**
   - 3.2. **GROUP** **[Constitutive]**
     - 3.2.1. HUMAN_GROUP **[Group]**
   - 3.3. **AMOUNT** **[Constitutive]**
4. **ENTITY** **[Top]**
   - 4.1 **CONCRETE_ENTITY** **[Entity]**
     - 4.1.1 LOCATION **[Concrete_entity]**
       - 4.1.1.1. *3_D_location* **[Location]**
       - 4.1.1.2. *Geopolitical_location* **[Location]**
       - 4.1.1.3. *Area* **[Location]**
       - 4.1.1.4. *Opening* **[Location | Agentive]**
       - 4.1.1.5. *Building* **[Location | Artifact$_{Agentive}$ | Telic]**
       - 4.1.1.6. *Artifactual_area* **[Location | Artifact$_{Agentive}$ | Telic]**

         recommended
     - 4.1.2. MATERIAL **[Concrete_entity | Telic]**
       - 4.1.3. ARTIFACT **[Concrete_entity | Agentive | Telic]**
         - 4.1.3.1. *Artifactual_material* **[Concrete_entity | Artifact$_{Agentive}$ | Material$_{Telic}$]**
         - 4.1.3.2. *Furniture* **[Concrete_entity | Artifact$_{Agentive}$ | Telic]**
         - 4.1.3.3. *Clothing* **[Concrete_entity | Artifact$_{Agentive}$ | Telic]**
         - 4.1.3.4. *Container* **[Concrete_entity | Artifact$_{Agentive}$| Telic]**
         - 4.1.3.5. *Artwork* **[Concrete_entity | Artifact$_{Agentive}$]**
         - 4.1.3.6. *Instrument* **[Concrete_entity | Artifact$_{Agentive}$ | Telic]**
         - 4.1.3.7. Money **[Concrete_entity | Artifact$_{Agentive}$ | Telic]**
         - 4.1.3.8. Vehicle **[Concrete_entity | Artifact$_{Agentive}$ | Telic]**
         - 4.1.3.9. Semiotic_artifact **[Concrete_entity | Artifact$_{Agentive}$| Telic]**
       - 4.1.4. FOOD **[Concrete_Entity| Telic]**

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

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

4.1.5. PHYSICAL_OBJECT **[Concrete_entity]0**

4.1.6. ORGANIC_OBJECT **[Concrete_entity]**

4.1.7. LIVING_ENTITY **[Concrete_entity]**

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

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

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

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

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

4.1.7.2.1. People **[Human]**

4.1.7.2.2. Role **[Human]**

4.1.7.2.2.1 Ideo **[Role]**

4.1.7.2.2.2 Kinship **[Role]**

4.1.7.2.2.3 Social_status **[Role]**

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

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

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

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

4.1.7.3.1. Plant **[Vegetal_entity]**

4.1.7.3.2. Flower **[Vegetal_entity]**

4.1.7.3.3. Fruit **[Vegetal_entity]**

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

4.1.8. SUBSTANCE **[Concrete_entity]**

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

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

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

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

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

4.2.1. QUALITY **[Property]**

4.2.2. PSYCH_PROPERTY **[Property]**

4.2.3. PHYSICAL_PROPERTY **[Property]**

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

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

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

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

4.3. **ABSTRACT ENTITY** **[Entity]**

4.3.1. DOMAIN **[Abstract_entity]**

4.3.2. TIME **[Abstract_entity]**

4.3.3. MORAL STANDARDS **[Abstract_entity]** *recommended*

4.3.4. COGNITIVE FACT **[Abstract_entity | Agentive]**

4.3.5. MOVEMENT OF THOUGHT **[Abstract_entity | Agentive]**

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

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

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

4.4.1. LANGUAGE **[Representation]**

4.4.2. SIGN **[Representation]**

4.4.3. INFORMATION **[Representation]**

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

4.4.5. UNIT OF MEASUREMENT **[Representation]**

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

4.5.1. PHENOMENON **[Event]**

4.5.1.1. *Weather verbs* **[Phenomenon]** *recommended*

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

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

4.5.2. ASPECTUAL **[Event]**

4.5.2.1. *Cause aspectual* **[Aspectual | $\text{Cause}_{\text{Agentive}}$]**

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

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

4.5.3.2. *Relational state* **[State]**

4.5.3.2.1. Identificational state **[Relational_state]** *recommended*

4.5.3.2.2. Constitutive state **[Relational_state]** *recommended*

4.5.3.2.3. Stative location **[Relational_state]** *recommended*

4.5.3.2.4. Stative possession **[Relational_state]** *recommended*

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

4.5.4.1. *Non relational act* **[Act]**

4.5.4.2. *Relational act* **[Act]**

4.5.4.2.1. Cooperative activity **[Relational_act | Agentive]** *recommended*

4.5.4.2.2. Purpose act **[Relational_act | Telic]** *recommended*

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

4.5.4.3.1 Caused motion **[Move | $\text{Cause}_{\text{Agentive}}$]**

4.5.4.4. *Cause act* **[Act | $\text{Cause}_{\text{Agentive}}$]**

4.5.4.5. *Speech act* **[Act]**

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

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

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

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

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

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

4.5.5. PSYCHOLOGICAL_EVENT **[Event]**

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

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

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

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

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

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

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

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

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

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

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

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

4.5.6.2.1. Transaction **[Change_possession]**

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

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

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

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

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

4.5.7.1.1. Cause_constitutive_change **[Cause_Relational_change]** *recommended*

4.5.7.1.2. Cause_change_of_state **[Cause_Relational_change]** *recommended*

4.5.7.1.3. Cause_change_of_value **[Cause_Relational_change]** *recommended*

4.5.7.2. *Cause_change_location* **[Cause_Change]**

4.5.7.3. *Cause_natural_transition* **[Cause_Change]**

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

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

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

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

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

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

General Ontology for Adjectives

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

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

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

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

   1.5. **Manner** **[Intensional]**

   1.6. **Object-related** **[Intensional]**

   1.7. **Emphasizer** **[Intensional]**

2. **EXTENSIONAL** **[Top]**

   2.1. **Physical_property** **[Extensional]**

   2.2. **Psychological_property** **[Extensional]**

   2.3. **Social_property** **[Extensional]**

   2.4. **Temporal_property** **[Extensional]**

   2.5. **Intensifying_property** **[Extensional]**

   2.6. **Relational_property** **[Extensional]**

Appendix B – EuroWordNet Top Ontology

Top[0]

1stOrderEntity[1]

Origin[0]
- Natural[21]
  - Living[30]
- Plant[18]
- Human[106]
- Creature[2]
- Anima[123]
- Artifact[144]

Form[0]
- Substance[32]
  - Solid[63]
  - Liquid[13]
  - Gas[1]
- Object1[62]

Composition[0]
- Part[86]
- Group[63]

Function[55]
- Vehicle[8]
- Representation[12]
  - MoneyRepresentation[10]
  - LanguageRepresentation[34]
  - ImageRepresentation[9]
- Software[4]
- Place[45]
- Occupation[23]
- Instrument[18]
- Garment[3]
- Furniture[6]
- Covering[8]
- Container[12]
- Comestible[32]
- Building[13]

2ndOrderEntity[0]

SituationType[6]
- Dynamic[134]
  - BoundedEvent[183]
  - UnboundedEvent[48]
- Static[28]
  - Property[61]
  - Relation[38]

SituationComponent[0]
- Cause[67]
  - Agentive[170]
  - Phenomenal[17]
  - Stimulating[25]
- Communication[50]
- Condition[62]
- Existence[27]
- Experience[43]
- Location[76]
- Manner[21]
- Mental[90]
- Modal[10]
- Physical[140]
- Possession[23]
- Purpose[137]
- Quantity[39]
- Social[102]
- Time[24]
- Usage[8]

3rdOrderEntity[33]

Appendix C – SIMPLE Extended Qualia Relations

Formal
isa
antonym_comp
antonym_grad
mult_opposition

Constitutive
made_of
is_a_follower_of
has_as_member
is_a_member_of
has_as_part
instrument
kinship
is_a_part_of
resulting_state
relates
uses
Property
causes
concerns
affects
constitutive_activity
contains
has_as_colour
has_as_effect
has_as_property
measured_by
measures
produces
produced_by
property_of
quantifies
related_to
successor_of
precedes
typical_of
contains
feeling
Location
is_in
lives_in
typical_location

Agentive
result_of
agentive_prog
agentive_cause
agentive_experience
caused_by
source

Artifactual_Agentive
created_by
derived_from

Telic
indirect_telic
purpose
Instrumental
used_for
used_as
used_by
used_against
Activity
is_the_activity_of
is_the_ability_of
is_the_habit_of
Direct Telic
object_of_the_activity

Derivational Relations

Derivation
AgentVerb
DeadjectivalNoun
DenominalAdjective
DenominalVerbNoun
Derivational
DeverbalAdjective
DeverbalNounVerb
EventVerb
InstrumentVerb
Nominalization
NounNoun
NounPropernoun
PatientVerb
ProcessVerb
StateVerb

Appendix D – EuroWordNet Semantic Relations

Relation Type	Parts of Speech	Labels	Data Types
NEAR_SYNONYM	N<>N, V<>V		Syn <>Syn
XPOS_NEAR_SYNONYM	N<>V, N<>AdjAdv, V<>AdjAdv		Syn <>Syn
HAS_HYPERONYM	N>N, V>V	dis, con	Syn <>Syn
HAS_HYPONYM	N>N, V>V	dis	Syn <>Syn
HAS_XPOS_HYPERONYM	N>V, N>AdjAdv, V>AdjAdv, V>N, AdjAdv>N, AdjAdv>V	dis, con	Syn <>Syn
HAS_XPOS_HYPONYM	N>V, N>AdjAdv, V>AdjAdv, V>N, AdjAdv>N, AdjAdv>V	dis	Syn <>Syn
HAS_HOLONYM	N>N	dis, con, rev, neg	Syn <>Syn
HAS_HOLO_PART	N>N	dis, con, rev, neg	Syn <>Syn
HAS_HOLO_MEMBER	N>N	dis, con, rev, neg	Syn <>Syn
HAS_HOLO_PORTION	N>N	dis, con, rev, neg	Syn <>Syn
HAS_HOLO_MADEOF	N>N	dis, con, rev, neg	Syn <>Syn
HAS_HOLO_LOCATION	N>N	dis, con, rev, neg	Syn <>Syn
HAS_MERONYM	N>N	dis, con, rev, neg	Syn <>Syn
HAS_MERO_PART	N>N	dis, con, rev, neg	Syn <>Syn
HAS_MERO_MEMBER	N>N	dis, con, rev, neg	Syn <>Syn
HAS_MERO_MADEOF	N>N	dis, con, rev, neg	Syn <>Syn
HAS_MERO_LOCATION	N>N	dis, con, rev, neg	Syn <>Syn
ANTONYM	N<>N, V<>V		Syn <>Syn
NEAR_ANTONYM	N<>N, V<>V		Syn <>Syn
XPOS_NEAR_ANTONYM	N<>V, N<>AdjAdv, V<>AdjAdv		Syn <>Syn
CAUSES	V>V, N>V, N>N, V>N, V>AdjAdv, N>AdjAdv	dis, con, non-f, rev , neg	Syn <>Syn
IS_CAUSED_BY	V>V, N>V, N>N, V>N, AdjAdv>V, AdjAdv>N	dis, con, non-f, rev, neg	Syn <>Syn
HAS_SUBEVENT	V>V, N>V, N>N, V>N	dis, con, rev, neg	Syn <>Syn
IS_SUBEVENT_OF	V>V, N>V, N>N, V>N	dis, con, rev, neg	Syn <>Syn
ROLE	N>V, N>N, AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn
ROLE_AGENT	N>V, N>N	dis, con, rev, neg	Syn <>Syn
ROLE_INSTRUMENT	N>V, N>N	dis, con, rev, neg	Syn <>Syn
ROLE_PATIENT	N>V, N>N	dis, con, rev, neg	Syn <>Syn
ROLE_LOCATION	N>V, N>N, AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn
ROLE_DIRECTION	N>V, N>N, AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn
ROLE_SOURCE _DIRECTION	N>V, N>N, AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn
ROLE_TARGET_DIRECTION	N>V, N>N, AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn
ROLE_RESULT	N>V, N>N	dis, con, rev, neg	Syn <>Syn
ROLE_MANNER	AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn
INVOLVED	V>N, N>N, V>AdjAdv, N>AdjAdv	dis, con, rev, neg	Syn <>Syn
INVOLVED_AGENT	V>N, N>N	dis, con, rev, neg	Syn <>Syn
INVOLVED_PATIENT	V>N, N>N	dis, con, rev, neg	Syn <>Syn
INVOLVED_INSTRUMENT	V>N, N>N	dis, con, rev, neg	Syn <>Syn
INVOLVED_LOCATION	V>N, N>N, V>AdjAdv, N>AdjAdv	dis, con, rev, neg	Syn <>Syn
INVOLVED_DIRECTION	V>N, N>N, V>AdjAdv, N>AdjAdv	dis, con, rev, neg	Syn <>Syn
INVOLVED_SOURCE _DIRECTION	V>N, N>N, V>AdjAdv, N>AdjAdv	dis, con, rev, neg	Syn <>Syn
INVOLVED_TARGET_DIRECTION	V>N, N>N, V>AdjAdv, N>AdjAdv	dis, con, rev, neg	Syn <>Syn
INVOLVED_RESULT	V>N, N>N	dis, con, rev, neg	Syn <>Syn
CO_ROLE	N>N	rev	Syn <>Syn
CO_AGENT_PATIENT	N>N	rev	Syn <>Syn
CO_AGENT_INSTRUMENT	N>N	rev	Syn <>Syn
CO_AGENT_RESULT	N>N	rev	Syn <>Syn
CO_PATIENT_AGENT	N>N	rev	Syn <>Syn
CO_PATIENT_INSTRUMENT	N>N	rev	Syn <>Syn
CO_PATIENT_RESULT	N>N	rev	Syn <>Syn
CO_INSTRUMENT_AGENT	N>N	rev	Syn <>Syn
CO_INSTRUMENT_ PATIENT	N>N	rev	Syn <>Syn
CO_INSTRUMENT_RESULT	N>N	rev	Syn <>Syn
CO_RESULT_AGENT	N>N	rev	Syn <>Syn
CO_RESULT_PATIENT	N>N	rev	Syn <>Syn
CO_RESULT_INSTRUMENT	N>N	rev	Syn <>Syn
IN_MANNER	V>AdjAdv, N>AdjAdv	dis, con, rev, neg	Syn <>Syn
MANNER_OF	AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn

Relation Type	Parts of Speech	Labels	Data Types
BE_IN_STATE	N>AdjAdv, V>AdjAdv	dis, con, rev, neg	Syn <>Syn
STATE_OF	AdjAdv>N, AdjAdv>V	dis, con, rev, neg	Syn <>Syn
FUZZYNYM	N<>N, V<>V		Syn <>Syn
XPOS_FUZZYNYM	N<>V, V<>AdjAdv, N<>AdjAdv		Syn <>Syn
IS_DERIVED_FROM	N, V, AdjAdv (across all)		VA<>VA
HAS_DERIVED	N, V, AdjAdv (across all)		VA<>VA
DERIVATION	N, V, AdjAdv (across all)		VA<>VA
ANTONYM	N<>N, V<>V, AdjAdv <> AdjAdv		VA<>VA
PERTAINS_TO	AdjAdv>N, AdjAdv>V		VA<>VA
IS_PERTAINED_TO	N>AdjAdv, V>AdjAdv		VA<>VA
HAS_INSTANCE	N>PN		Syn>I
BELONGS_TO_CLASS	PN>N		I>Syn