Towards Best Practice for Multiword Expressions in Computational Lexicons

Nicoletta Calzolari[1], Charles J. Fillmore[2], Ralph Grishman[4], Nancy Ide[3], Alessandro Lenci[1], Catherine MacLeod[4], Antonio Zampolli[1]

[1]Istituto di Linguistica Computazionale, CNR, Pisa
Università di Pisa, Dipartimento di Linguistica
[2]ICSI, University of California Berkeley
[3]Department of Computer Science, Vassar College
[4]New York University

Abstract

The importance and role of multi-word expressions (MWE) in the description and processing of natural language has been long recognized. However, multi-word information has often been relegated to the marginal role of idiosyncratic lexical information. The need for MWE lexicons grows even more acute for multi-lingual applications, for which (sometimes complex) correspondences must be identified, classified, and recorded. Within the XMELLT and ISLE projects we have started to investigate the potential to develop multi-lingual, multi-word expression lexicons incorporating both syntactic and semantic information. We aim at specifying means to acquire and represent multi-word lexical entries for multiple languages, and establishing uniform (or inter-translatable) standards for describing multi-word lexical entries. We explored theoretical approaches used in large lexicon-building projects, in particular FrameNet and SIMPLE. They constitute interesting frameworks for the explicit syntactic and semantic representation of MWEs, due mainly to their ability to capture semantic multidimensionality, through frame elements and qualia relations respectively. We also developed an abstract data model for lexical information together with a representation in XML for it. Our goal is to define a set of *minimal lexicon "objects"*, which can serve not only as a model for MWEs but also for lexical data in general.

1. Introduction

The importance and role of multi-word expressions (MWE) in the description and processing of natural language has been long recognized. However, despite the fact that large computational lexicons have begun to exist that contain both syntactic and semantic information, multi-word information has often been relegated to the marginal role of idiosyncratic lexical information, or has been addressed in terms of specific types of word combinations only. The need for MWE lexicons grows even more acute for multi-lingual applications, for which (sometimes complex) correspondences must be identified, classified, and recorded.

Recognizing this, we undertook a one-year NSF-funded pilot project (XMELLT[1]) to investigate the potential to develop multi-lingual, multi-word expression lexicons incorporating both syntactic and semantic information. The project has as its goals to (1) specify means to acquire and represent multi-word lexical entries for multiple languages; (2) create a small number of multi-word entries for support verbs and noun compounds; and (3) establish uniform (or inter-translatable) standards for describing multi-word lexical entries at the levels of syntax and morpho-syntax and lexical semantics. XMELLT involved four U.S. institutions: ICSI (Berkeley), Vassar College, New Mexico State University, and New York University. The Istituto di Linguistica Computazionale of the Italian National Research Council also participated, in the context of the EAGLES-ISLE[2] standardisation project (Calzolari et al., 2002), within the Multilingual Lexicon Working Group.

2. What is an MWE?

In different theoretical or practical contexts the term *multiword expression* (MWE) is used to describe different but related phenomena, including fixed or semi-fixed phrases, compounds, support verbs, idioms, phrasal verbs, collocations, etc. At the level of greatest generality, all of these phenomena can be described as *a sequence of words that acts as a single unit at some level of linguistic analysis*. In addition, they exhibit some or all of the following behaviors:

1. reduced syntactic and semantic transparency;
2. reduced or lack of compositionality;
3. more or less frozen or fixed status;
4. possible violation of some otherwise general syntactic patterns or rules;
5. a high degree of lexicalization (depending on pragmatic factors);
6. a high degree of conventionality.

MWEs differ in the degree to which the features (1)-(6) occur, and therefore span a continuum from full-fledged compositional and productive constructions to collocations to fixed idioms. MWEs can be regarded as lying at the interface between grammar and lexicon. In fact, they are usually instances of well productive syntactic patterns which nevertheless exhibit a peculiar lexical behavior. As a result, MWEs defy naïve attempts to establish a border between grammar and lexicon in terms of the opposition between rule productivity and lexical idiosyncrasy.

This is one of the causes of the difficulties that MWEs raise both under the theoretical and the computational point of view:

[1] "Cross-lingual Multiword Expression Lexicons for Language Technology", Nancy Ide, Vassar, PI; NSF Award No. 9982069, May 1,2000 – December 31, 2001.

[2] ISLE (*International Standards for Language Engineering*) is a transatlantic standards oriented initiative under the Human Language Technology (HLT) programme within the EU-US (EC-NSF) International Research Co-operation. It is a continuation of the European EAGLES (*Expert Advisory Group for Language Engineering Standards)* initiative funded by the European Commission since 1993.

- it is difficult to provide clear-cut boundaries to the domain of MWEs;
- from the computational point of view, the identification of MWEs is a traditional hot-topic and an essential ingredient in whatever NLP task, from efficient CLIR and IE to MT and text generation, but are still missing in most computational lexicons of reasonable size;
- from the multilingual perspective, often MWEs in a certain source language do not find direct lexical equivalents in another, and again they suggest highly complex lexicon-to-lexicon and lexicon-to-grammar interactions;
- they concern both general and terminological lexicons.

While collocations and idioms might appear in a multi-word expression lexicon, in the XMELLT project and now in ISLE we focus on multi-word expressions that are, on the one hand, productive, and, on the other, demonstrate regularities (e.g. in argument structure) that can be generalized to classes of words with similar properties. In particular, we are concerned with devices in grammar that allow for the production/analysis of new MWEs rather than learned phrases; as such, the information we are concerned with is at the intersection of grammar and lexicon. A primary motivation for this approach is to enable the fully or semi-automatic recognition/acquisition of MWE lexicon entries.

We studied in depth two MWE phenomena: *support verbs* (or *light verbs*) and *noun compounds* (or complex nominals). Because both of these phenomena lie at the center of the variation specter in compositionality that can be observed in MWEs, exhibiting internal cohesion together with a high degree of variability in lexicalization and language-dependent variation, they represent the hard cases among MWEs. As such, their representation demands tackling core issues for the representation and status of MWEs in multilingual lexicons. Henceforth, with the term MWE we will only refer to complex nominals and to support verb constructions.

The paper is organized along a number of crucial issues that we regard as of the utmost importance in the domain of MWEs, and which provide a potential roadmap for the ISLE-XMELLT proposals on this topic. We will provide examples and analyses of all these, together with a preliminary set of related best practice recommendations for the treatment of MWE in mono- and multi-lingual computational lexicons.

2.1. Establishing the boundaries for the investigation

For both support verbs and complex nominals, it is possible to choose between a narrow and a wide definition. The narrow definition typically restricts the dimension and the pervasiveness of the phenomena, and leads towards purely lexicalist analysis. Narrow characterisations of the topic are however not very satisfactory, and therefore an enlargement of the basic original notion is often needed.

2.1.1. Support verbs

Narrow definition: support verbs should be restricted to the case of semantically light verbs (e.g. *have a coffee, take a shower, make a call, make an acquisition*, etc.)

However, the definition of light verb is problematic (especially if light verb means a verb which brings no semantic contribution other than the one of turning a noun into a verbal expression). Support verbs actually represent a much wider phenomenon:

make an acquisition1
complete an acquisition1
undertake an acquisition1
take action1 (e.g. military action)
bring action2 (e.g. legal action)
make an application1
have an application1 in
decide on an application1 (consider, hear)
get an application1 (receive, take)
submit an application1 (file)

In a still narrow sense, support verb constructions include syntactic patterns formed by V + deverbal N. On the other hand, this definition should also be enlarged to event/result/abstract nouns not necessarily morphologically derived:

dare un ceffone (to slap)
provare rancore (to bear sb a grudge)
fare una festa (to have a party)
fare festa (to have a holiday)
fare festa a qno (to give sb a warm welcome)
prestare attenzione (to pay attention)
fare la guerra (to wage war)

These examples do not contain deverbal nouns, but still behave in a quite similar way to 'regular' constructions with deverbals.

We can distinguish two types of support verbs:

- type 1 - when they combine with an event noun (either deverbal or not), these verbs present their subjects as participants in the event most closely identified with the noun:

take an exam, give an exam
perform an operation, undergo an operation
ask a question
make a promise

The type 1 support verbs mainly correspond to some of the Mel'cuk (Mel'cuk & Polguère, 1987) lexical functions (e.g. Oper1 and Oper2).

- type 2 - the subject of these verbs belongs to some scenario associated with the full understanding of the event type designated by the noun:

pass an exam, fail an exam, grade (evaluate) an exam
survive an operation
answer a question
keep a promise

The type 2 support verbs deal with some implications of the basic event, but are not a part of the event itself. For instance, in the case of *grade an exam*, the subject participates in the scenario implied by the exam event, which includes besides the people giving the exam also the individuals who have to evaluate it.

XMELLT has adopted an operative definition of support verbs which is roughly equivalent to type 1 above: V+N constructions expressing the same event as the event named by the noun, and whose subject participates in this event. Thus, for the noun *operation*, we included the

verbs *perform* and *undergo,* but not *observe* or *remember*.[3]

We collected English support verbs for 50 deverbal nouns in NOMLEX[4] (MacLeod, *et al. 1998*), recording argument structure and some limited semantic features, using a subset of Mel'cuk's lexical functions. We then encoded support verbs for Italian, including both (1) nominalizations that did not correspond to the NOMLEX entries, in order to assess the applicability of the lexical functions to Italian support verb constructions; and (2) translations of the English support-verb constructions, whether the Italian equivalents were MWEs or not, to be able to determine the types of correspondences that can hold between lexicons in the different languages.

We have extracted many examples from an Italian corpus, providing evidence to the lack of regularity in the correspondence between 'lexical function' and choice of support verb. There are, as known, many lexical ways to express the same lexical function, many idiosyncracies, and often the default typical verb for a specific function is not attested at all. Also words belonging to the same semantic class may preferentially select different verbs to express similar meanings. This makes it essential to devise a strategy of i) acquiring this information from text corpora, and ii) listing in a computational lexicon at least those support verbs which are lexically dedicated to/selected by a specific noun. Point (i) can be done rather staightforwardly, possibly in a multilingual environment, starting from typical verbs accompanying nominalizations, categorizing and documenting lexical functions, and creating equivalence classes of verbs associated to different lexical functions.

2.1.2. Complex nominals

XMELLT also considered complex nominals, in order to describe the relationships exhibited in the compounds using a descriptive framework that is a compromise of notions from frame-semantics, qualia theory, and the syntactic relations employed in the NOMLEX project.

According to a traditional definition, in English compounds have the stress placed on the first element (e.g. *blackbird*). However, this represents a too narrow criterion to identify compounds, especially once the needs of computational systems are taken into consideration. Complex nominals in English usually appear with a N head pre-modified by a N, adjective, possessive phrase or gerund:

food container
butcher's knife
censorship controversy
health crisis
hunting dog
environmental risk

Language variation is high. For instance, in French and Italian, complex nominals correspond to a N head plus a post-modifier element, usually:

N + Adjective
N + PP
N + Vinf

Obviously there is no one-to-one correspondence between syntactic patterns in the two languages:

*coltello **da** macellaio*	*butcher's knife*
*carta **di** credito*	*credit card*
*carta telefon**ica***	*phone card*
*agenzia **di viaggi***	*travel agency*
*film **per** adulti*	*adult movie*
*macchina **da** scriv**ere***	*typewriter*

We examined a wide variety of compounding types, including (a) nominalizations (event nominalizations, result nominalizations, etc); (b) artifact names (where the components can be analyzed in terms of such relations as part-whole, cause-effect, substance-item, function-item, occasion-item, and the like); (c) words from natural taxonomies (where the relations will involve subtype naming, part-whole, organism-habitat/range, etc.); (d) general situation-labeling words (situation, condition, event, etc.). We considered these phenomena mainly in respect to their functioning as the heads of compounds, but also, where relevant, as modifiers. Thus, for *bus* we include not only *school bus* but also *bus ticket*: the latter makes use not only of frame information connected with the head noun, *ticket*, but also information connected with the *bus* frame, having to do with becoming an authorized passenger on a bus. There is often a mutual dependency between head and dependent noun, and it may happen that the 'semantic head' does not correspond to the syntactic head.

The modifiers of the English target nouns that we examined include not only words of category singular-noun-stem, but also possessives (*butcher's knife*, *men's room*), joined nouns modifying nouns that stand for relationships or interactions (*army-navy game*, *parent-child disagreements*), plural nouns (*damages verdict*), and relational (as opposed to descriptive) adjectives (*educational policy, philosophical society*). The adjectives are of the type referred to as pertainyms in WordNet, and usually have similar functions to nouns in the same location. We explored in particular the possibility of discovering semantic classes of nouns capable of occurring as modifiers, and listing their members. Thus, in the case of the *ticket* example, we include names of public conveyances (*train*, *airplane*, *bus*, *shuttle*, etc.), names of entertainments (*theater*, *opera*, *museum*, *concert*, etc.; *football*, *baseball*, *hockey*, etc.).

2.2. Problematic Issues

The main difficulty is that both support verbs constructions and complex nominals instantiate well-formed syntactic patterns that also correspond to something that is usually not classified as a multiword, or is only loosely described as such, e.g. *:*

give a speech	vs.	*listen to a speech*
take a decision	vs.	*change a decision*
fare un invito	vs.	*accettare/rifiutare un invito,*
carta telefonica	vs.	*conversazione telefonica*
(lit. phone card)		(lit. phone conversation)
discussione da salotto	vs.	*discussione da Giovanni*
(lit. society gossip)		(lit. discussion at John' s place)

Moreover, the syntactic pattern – as said above - is not totally predictable:

travel agency	*agenzia di viaggi*
real estate agency	*agenzia immobiliare*

[3] Thus, while it is possible to observe or remember one' s own medical operation, that possibility is not part of the semantics of the noun *operation.* By contrast, the verbs *perform* and *undergo* explicitly relate their subject NPs to specific roles in the noun related event.

[4] http://www.cs.nyu.edu/cs/projects/proteus/nomlex/

wedding agency *agenzia matrimoniale*

Finally what is a MWE in a language might not be a MWE in another language (e.g. *cucchiaino da caffé* lit. "coffee spoon" vs. *tea spoon*, whose Italian equivalent *cucchiaino da tè* could hardly be regarded as a MWE), or might be translated in a different way (e.g. *pay attention* → *prestare attenzione* lit. "lend attention"). This implies, e.g. for machine translation, some conceptual representation. The 'encoding' process must find and appropriate MWE in L2 if it is called for: this is analogous to "blocking/pre-emption", where a regular compositional process is not carried out (dispreferred) because the semantic space occupied by the concept associated with that formation is already claimed by some ready-made expression.

2.3. Linguistic diagnostics for MWEs

It is necessary to single out some diagnostics that may help us to identify which linguistic expressions belong to MWEs. We identified several key features of complex nominals, e.g.:

- The modifier in a complex nominal has reduced or no *semantic referentiality;*

(1) * *Gianni ha comprato un comprato una* ***bottiglia da vino*** *che berrà a cena.*

'John has bought a wine bottle, which he will drink at dinner'

(2) *Gianni ha comprato una* ***bottiglia di vino*** *che berrà a cena.*

'John has bought a bottle of wine, which he will drink at dinner'

In *bottiglia* ***da*** *vino, vino* is somehow referentially non-accessible or incorporated into the larger MWE, so that it cannot be drunk. Conversely, in *bottiglia* ***di*** *vino, vino* is not incorporated and perfectly accessible from the outside. Similarly something can be a *hunting knife*, even though it has never been used to go hunting, etc.

- Truly complex nominals define a new subtype of the entity denoted by the head of the compound:

il tavolo da giardino (garden table) → is a particular subtype of tables

il tavolo per ***il*** *giardino* (the table for the garden) → is not a new type of table, and the modifier specifies the particular function that a given table may happen to have in a certain context.

- Particular syntactic clues may help to identify patterns that qualify as complex nominals. In Italian for instance complex nouns often occur without a determiner in the postmodifier PP:

carta di credito		
negozio di scarpe	vs.	*negozio delle scarpe*
computer da tavolo	vs.	*computer sul tavolo*
oggetto di studio	vs.	*oggetto dello studio*

It is important to stress the fact that these tests, although useful, are never completely discriminating nor decisive. At the very end, it is not even necessary to have clear-cut tests, since the main point is to be able to provide a satisfactory explicit characterization of the internal structure of these constructions.

3. Representation

Support verbs and complex nominals are similar in being generally instances of well-productive and attested syntactic patterns (e.g. V+NP or N+PP), and yet they show various degrees of lexicalization. On the other hand a purely, and 'brute-force', lexicalist approach is not enough, especially if this is intended as merely listing MWEs in a lexicon. This point is supported by some general motivations:

- we loose generalizations;
- we loose the possibility to produce a proper interpretation of these constructions;
- we run into problems when operating in a multilingual environment, when something that is a MWE in a certain language has to be expressed in the target language in terms of a normal syntactic pattern.

As a first approximation, we can regard complex nominals and support verbs as *lexical constructions.* They have lexical-like behaviors (both at the semantic and the syntactic level), and yet their interpretation depends heavily on the particular relation that holds between the components of the MWE (this is for instance a major difference with respect to fully idiomatic and frozen expressions, whose interpretation turns out to be totally detached from the interpretation of the sub-parts). Some examples displaying both syntactic and semantic variation, and the well-known lack of correspondence between the two:

hunting knife → a knife *used for* hunting
negozio di scarpe → a shop *that sells* shoes
stringa da scarpe → a lace *for* shoes
lavaggio a mano → washing *by using* hands
mal di macchina → sickness *caused by* the car
vestito da sera → dress *to be used in* the evening
vestito da matrimonio → dress *designed for* weddings

Despite the high degree of variation, the semantic relations between the constituents are a function of the interaction of the semantics of the two elements involved. Moreover, the interpretation of MWEs is similar to the one of regular syntactic patterns. In most cases MWEs are organized in sorts of *semantic paradigms* or *variation classes*, which depend on the particular semantic functions the elements of the MWEs fulfill, and which are syntactically realized in an often quite predictive and constant way (cf. the distribution of prepositions in Italian in different semantic types of compounds, in Busa & Johnston, 1996).

We have therefore concentrated our interest i) on some theoretical approaches to the lexicon, which allow for a representation of the relational nature of lexical items, such as Frame Semantics, Generative Lexicon, Lexical Functions, Lexical Conceptual Structure, and ii) on existing and available lexicons already providing the basic resources (core notions, formal apparatus, and general schema to characterize the problem) to represent the internal constitution of MWEs, such as FrameNet and SIMPLE. As we will show below, FrameNet and SIMPLE make appeal to specific frame structures and qualia relations of the head noun respectively, together with the semantic type of the modifier, to account for the underlying semantic motivation in MWE. It will be necessary to undertake a detailed corpus analysis to determine from corpus attestations which frame elements or qualia can get instantiated as a modifier word, and how

they are realised morpho-syntactically. In a multilingual context, looking for the frame/qualia structure of the head noun, and to its interaction with the semantic type of the modifier, is more significant than trying to find some general-purpose classification of modification types.

3.1. Recommendations and standards

Because we are creating MWE entries across a variety of languages, it is essential not only to adopt the same model for different languages, but also to determine the types of links (constraints, conditions, etc.) that should be made among entries in different mono-lingual lexicons in order to facilitate translation. With respect to this issue, the relation with ongoing standardization initiatives in computational lexicography has a crucial role in XMELLT.

It is often argued that in terms of lexicon building, MWEs raise the question of how to decide between i) listing complex forms in the lexicon, and ii) writing rules for deriving them. However, in the context of designing a multilingual lexicon, this distinction is not always relevant, because the matching of equivalents across languages does not necessarily involve matching MWEs with MWEs and free forms with free forms[5]. There are regularities (or at least tendencies) in each language, but they don't match. Thus a multilingual lexicon for MWEs must meet the following criteria. It must both *describe* and *list*:

- describe the syntactic behavior of the MWE constructions, with particular attention to syntactic peculiarities and morpho-syntactic constraints (e.g., lack of determiner, necessary occurrence in the plural, etc.);
- relate the MWE to the normal productive syntactic patterns underlying it, where relevant;
- characterize the semantics of each constituent and the manner of their composition (e.g. looking at the semantic structure of the head noun and at the variety of modifiers it can select by virtue of its meaning, and at how the semantics of the two nouns interact when they co-occur);
- identify and make explicit the semantic relations between the constituents, and identify general semantic paradigms;
- list the particular instantiations of a certain semantic paradigm;
- calibrate the different degrees of lexicalization.

Satisfying these criteria for a proper representation and description of MWEs demands a *structural approach* to the lexicon, where:

- particular attention is devoted to the relational character of lexical items;
- it is possible to have access to the *semantic constitution* of lexical items.

Moreover, it is necessary to list idiosyncratic behaviors, and to this purpose it an interesting strategy may be to link a lexicon to very large "dassified" repositories of textual co-occurrences, i.e. collocational/syntagmatic data, especially for multilingual applications.

[5] The term **free form** here is used to refer to single words or semantically transparent phrases.

As already mentioned, close interactions have occurred between XMELLT and the ISLE computational Lexicon Working Group (CLWG). This has mostly concerned the issue of proposing a standard model for the representation of MWEs. The following are some possible steps to achieve this goal:

a. select some basic notions and formal resources (taking them from existing computational lexicons and frameworks) which are necessary to represent the internal semantic constitution of MWEs, as well as of their components;
b. adopt a layered approach to the representation of the MWEs,to be incorporated within MILE (Multilingual ISLE Lexical Entry), representing the prospective output of the ISLE CLWG recommendations. MILE would actually provide a series of levels of representation for MWEs, at various degrees of generalization, which could be specifically targeted or used by different NLP applications:
    1. MWEs as syntactically complex unit, but semantically simple;
    2. representation of the internal relational structure of MWEs
    3. representation of the semantic paradigms which are instantiated by the MWE' s
c. allow for a representation of the lexical idiosyncrasies shown by the MWEs (useful for text generation).

It is essential that MILE will offer a multi-layered encoding of the MWEs, so that the user will be able to choose the suitable level of granularity for their description in the lexicon.

3.2. Available representational devices

To satisfy point (a) we explored theoretical approaches used in several large lexicon-building projects that allow for a representation of relations among lexical items, in particular FrameNet, based on **frame semantics** (Fillmore & Baker, 2001), and the SIMPLE lexicons (Lenci et al., 2000) based on the **generative lexicon** (Pustejovsky, 1995). We considered the representation for the support verb entries in both English and Italian using the categories and relations defined within these schemes, and determined that they provide us with the formal apparatus which is needed to describe both syntactically and semantically the internal constitution of MWEs.

Both FrameNet and SIMPLE represent interesting frameworks for the explicit syntactic and semantic representation of complex nominals, due in particular to their ability to capture semantic multidimensionality, through frame elements and qualia relations respectively. For instance, for MWE whose head is *container*, they can represent, among others, the following meaning dimensions:

a. FrameNet

Container Frame:
Frame Elements: Material, Contents, Size, and Function.
Material:
aluminum container, glass container, metal container, tin container
Contents:

food container, beverage container, trash container, water container, milk container, fuel container

`Size:`

3 quart container

`Function:`

shipping container, storage container

b. SIMPLE

Qualia Relations for "containers" compounds:

`Constitutive:` *made_of* `([MATERIAL])`

aluminum container, glass container, metal container, tin container

`Telic:` *contains* `([ENTITY])`

food container, beverage container, trash container, water container, milk container, fuel container

`Constitutive:` *size* `([QUANTITY])`

3 quart container

`Telic:` *is_used_for* `([EVENT])`

shipping container, storage container

Both these frameworks contain the formal apparatus allowing lexicon developers to describe the internal semantic constitution of the elements composing the complex nominals. In Italian, an additional difficulty is the fact that the association between a preposition and a qualia relation is not straightforward. There are tendencies which can be captured in the form of preferences, but these correspondences seem to involve not only the qualia relations, but also the semantic types of the two nouns. Here a few examples of a possible simplified representation in SIMPLE:

coltello da macellaio (butcher's knife)◗TELIC(used_by)Y
[Human] ◗ PPda

coltello di plastica (plastic knife) ◗ CONST(made_of) X
[Material] ◗PPdi

coltello da tavola (table knife) ◗ TELIC (used_in) Z
[Location] ◗PPda

coltello da caccia(hunting knife)◗TELIC(used_in_activ.)E
[Activity] ◗Ppda

piatto di legno (wooden dish)◗ CONST (made_of) X
[Material] ◗PPdi

piatto di pasta (dish of pasta)◗ CONST (contains) X
[Food] ◗PPdi

In fact, the interpretation of compound nouns, crucially depends on the multidimensional semantic structure of the head and the modifier, as well as on phenomena of coercion and co-composition occurring between them. It is interesting the fact that the interpretation of MWEs can be done using the same representational devices already available in both lexicons for interpreting regular noun constructions: MWEs and regular noun constructions seem to share and make appeal to the same general principles of semantic constitution of lexical items and their combinatorics, e.g. in terms of frame/qualia structures.

3.3. Abstract data model for lexical information

At the same time we developed an abstract data model for lexical information (Ide *et al.*, 2000) together with a representation in XML for it. The support verb entries in English (originally represented in the NOMLEX format) and the Italian entries (originally represented in the SIMPLE format) were then mapped onto the XML representation, in order to render them in a common format and to enable linkage. We also developed scripts using the XML Transformation Language (XSLT[6]) to extract specified pieces of the entries and display them in a readable format on a web browser. This small experiment yielded several interesting results: first, it served as a proof of concept that our abstract model for lexical information is powerful enough to represent the required categories and relations. Second, it revealed several problems with the original formats. Finally, it demonstrated that lexical resources in very diverse formats can be mapped to a common format. This is a fundamental criterion underlying the design of the abstract model and its XML instantiation: lexicon developers should be able to use internally or specially developed formats for their data, and it should be trivial to map those formats to a more abstract model without information loss, for purposes of merging, comparison, exchange, etc.

We are developing an approach to lexicon representation that is object-based: objects in the model are various pieces of a lexical entry. We propose to use the GENELEX/SIMPLE model as a starting point, and to extend it as needed to accommodate MWEs and their linkage cross-lingually. Our goal – which is also one of the main goals of ISLE for the definition of the MILE - is to define a set of *minimal lexicon "objects"* and specify fully the ontological relations and other relations among them, which can serve not only as a model for MWEs but also for lexical data in general. The central component of this activity is the development of an ontological description of the structure of lexical entries, focusing on the identification and the formal definition of hierarchies of "lexical constructions". These will form the basic structures to build templates of lexical entries. Inheritance relations among these structures must also be identified, as well as the possible constraints acting on them, all of which can be formalized using the Resource Definition Framework (RDF). This object-oriented approach to lexical architecture enables setting up a particularly expressive framework, particularly suited to capture various types of linguistic generalizations in MWEs. Cross-linguistically, MWEs distribute along several equivalence classes or paradigms, which can be adequately described only by taking into account highly integrated morphological, syntactic, and semantic information.

The ontological description of lexicon structure, inheritance relations, and constraints can be formalized as a set of RDF schemas augmented by the extensions defined in DAML+OIL[7]. This will ultimately enable exploiting powerful semantic web technology to access lexical information by devising a "layered" XML/RDF specification for lexical entries.

Obviously, when representing MWEs in computational lexicons we want to exploit regularities in forming MWEs and their translations to avoid simply listing translation equivalents, as well as to account for "new" MWEs that follow regular rules of formation. So, to provide a simple example, in Italian, *butcher' s knife* is *coltello* ***da*** *macellaio*, and *steel knife* is *coltello* ***di*** *acciaio*. Our goal is to find means to represent this type of phenomenon to enable identification of the appropriate translation for, say,

[6] http://www.w3.org/TR/xslt/

[7] http://www.daml.org

chef's knife vs. *plastic knife*, which follow the same pattern in the Italian translation, without explicitly listing the correspondences for each possible noun pair. To accomplish this, some mechanism is needed to associate a noun appearing in the role of "typical user" with the *da* construction, and to associate the *di* construction with a noun in the role "made-of", so that the appropriate equivalent can be identified/generated. Because such phenomena are specific to particular language pairs, this information should not be included in the mono-lingual lexicons, but rather should be associated with the link between the appropriate entities in the lexicons itself. This is where semantic web technologies can be exploited to associate pre-defined processes with lexical information both within and between lexicons that can not only exploit inferencing capabilities, but also dynamically construct corresponding entries using associated rules—or even rules (processes) which are themselves "constructed" on-the-fly, on the basis of available information.

4. References

Busa, F., Johnston, M. (1996). "Cross-linguistic semantics for complex nominals in the generative lexicon". In *Proceedings of the AISB Workshop on Multilinguality in the Lexicon*, Brighton, University of Sussex.

Calzolari, N., Zampolli, A., Lenci, A. 2002. "Towards a Standard for a Multilingual Lexical Entry: the EAGLES/ISLE Initiative". In A. Gelbukh (ed.), *CICLing-2002 Third International Conference on Intelligent text processing and Computational Linguistics*, Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg New York.

Fillmore, C. J., Baker, C. F. (2001). "Frame Semantics for Text Understanding", in *Proceedings of WordNet and Other Lexical Resources Workshop*, NAACL, Pittsburgh, June, 2001.

Ide, N., Kilgarriff, A., Romary, L. (2000). A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000*, Stuttgart, 113-126.

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

MacLeod, C., Grishman, R., Meyers, A., Barrett, L., Reeves, R., (1998), "NOMLEX: a Lexicon of Nominalizations", Proceedings of EURALEX' 98, Liege, Belgium.

Mel'cuk, I. , Polguère, A. (1987). "A forlam lexicon in the Meaning-Text Theory (or How to Do Lexica with Words)". In *Computational Linguistics*, 13.

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