Semi-Automatic Derivation of a French Lexicon from CLIPS

Nilda Ruimy[#], Pierrette Bouillon*, Bruno Cartoni*

[#] Istituto di Linguistica Computazionale - CNR – Pisa, Italy
* TIM/ISSCO, ETI, Université de Genève
nilda.ruimy@ilc.cnr.it
pierrette.bouillon@issco.unige.ch
bruno.cartoni@eti.unige.ch

Abstract

In this paper we describe the methodology developed in the framework of a feasibility study for the derivation of a semantically annotated French lexicon from a monolingual Italian lexical resource. Firstly, an outline of the source lexicon is provided. Then, the two different and complementary strategies that have been experimented for pairing off the relevant monolingual Italian entries and their translational equivalents are described. Finally, the results achieved through each of the illustrated methodologies are presented, their viability is evaluated and a general assessment of the experiment performed is provided.

1. Introduction

Building large-scale computational lexicons from scratch has been indisputably recognized as a quite challenging, costly and time-consuming task. We therefore decided to investigate the feasibility of creating a semantically annotated French lexicon, by deriving the core semantic properties to be assigned to its entries from an Italian computational lexical database richly endowed with outstanding information. To derive the French entries, a crucial step consists in correctly pairing off the French word senses with the corresponding semantic units of the Italian lexicon from which the lexical information will be borrowed. In this paper, the two different and complementary approaches taken in this regard are described, preliminary results are presented and the viability of the whole methodology is assessed.

2. The Monolingual Lexical Database

Our monolingual source of information is the largest computational lexical knowledge base of Italian language. CLIPS (Ruimy et al., 2002) is a multi-level, general-purpose lexicon in which words are encoded at four different levels of linguistic description. The whole resource consists of 55,000 lemmas described at the phonological, morphological and syntactic level and 55,000 word senses encoded at the semantic level, all in accordance with the international standards set out in the PAROLE-SIMPLE model (Ruimy et al, 1998; Lenci et al., 2000). As a matter of fact, CLIPS builds on and extends the Italian version of the twelve PAROLE-SIMPLE European lexicons[1] that share a common theoretical model, representation language and building methodology.

The linguistic description of entries offers very fine-grained information, most relevant for NLP applications. In a CLIPS entry, all the phonological, morphological and inherent syntactic properties of a headword are represented. Its subcategorization pattern is/are described in terms of optionality, syntactic function, syntagmatic realization and morpho-syntactic, syntactic and lexical properties or constraints of each slot filler.

Following the SIMPLE model, the theoretical approach to the content and representation of information at the semantic level is essentially grounded on a revisited version of some fundamental aspects of J. Pustejovsky's Generative Lexicon Theory (1995,1998). A CLIPS semantic unit is richly endowed with a wide set of fine-grained, structured information. First among them, the sortal classification: the lexicon is in fact structured according to a multidimensional type hierarchy based on both hierarchical and non-hierarchical conceptual relations, taking into account the principle of orthogonal inheritance (Pustejovsky & Boguraev, 1993). Other relevant information types in a word entry are its domain of use; type of denoted event; synonymy and morphological derivation relations; membership in a class of regular polysemy as well as any relevant distinctive semantic features. Particularly outstanding is the information encoded in the *Extended Qualia Structure* and the *Predicative Representation*. The *Extended Qualia Structure* allows modelling both the different meaning dimensions of a word sense and its relationships to other lexical units, by means of 56 semantic relations subsumed by the original Qualia Structure's four roles[2] (Pustejovsky, 1995). These semantic relations - which make it possible to characterize a word sense (hypernymic relation), describe its meronymic properties, indicate its origin and its function - link either intracategorial or intercategorial semantic units. As to the *Predicative Representation*, it describes the semantic scenario the word sense at hand is involved in and characterizes its participants in terms of thematic roles and semantic constraints. Moreover, in a word's description, syntactic and semantic information are related to each other through the projection of the predicate-argument structure onto its syntactic realization(s).

Considering the wealth and fine-grainedness of the lexical knowledge that CLIPS offers, its exploitation for deriving other lexical resources seems quite advisable. Hence the interest of devising tools for reliably detecting the lexical entries whose information is of particular interest.

[1] built in the framework of the European projects PAROLE and SIMPLE.

[2] The formal, constitutive, agentive and telic roles.

3. The Feasibility Study

To assess the feasibility of deriving a semantically annotated French lexicon using CLIPS lexical knowledge, a study was conducted whereby a twofold approach was adopted for relating French word senses to the corresponding CLIPS semantic units. On the one hand, for constructed words, the *cognateness* of a set of Italian and French suffixes was exploited. On the other hand, an approach based on *sense indicators* was taken up for all cases in which the first methodology could not be applied.

3.1. The Cognate Approach

The first strategy is based on the fact that French and Italian languages share a lot of similarities in terms of lexical structure and syntactic information. Since they both derive from a common root language (Latin), their core lexicon is quite similar (Geysen, 1990). In particular, research has demonstrated that their morphological systems show an important parallelism (Namer, 2001). The strategy proposed here takes advantage of this similarity and is guided by the two following hypotheses: (1) morphologically constructed words usually have sense(s) that are largely predictable from their structure and (2) Italian suffixed items have one (or more) equivalent(s) constructed with the corresponding French suffix that cover(s) all the senses of the Italian word.

3.2.1. Methodology

More concretely, the cognate strategy can be summarized as follows. The basic resource is a bilingual dictionary from which we extracted Italian constructed headwords and, for each of their senses, the different possible translations, for example: *torrefazione*, (1) *torréfaction*, (2) *maison du café*. We then assumed that if an Italian word encoded in CLIPS has, in our bilingual database, the same translation for all its senses, this French equivalent will share with the Italian word all the CLIPS entries. For example, *villagio* has two senses in CLIPS, one for *the place* and one for *the group of people living in a village.* Since this word is always translated with *village*, we can infer, **with no further analysis**, that *village* shares the CLIPS entries of *villagio*. In the following, we evaluate this approach with three suffixes: *-tà, -zione* and *–aggio.*

3.2.2. Experimental Data

To evaluate the relevance of this method, we extracted randomly from CLIPS 79 words ending in *–aggio*, 80 in *–tà* and 56 in *–zione*. For each of these words, we then checked in the Robert-Signorelli[3] the number of different translations (see table 1) and, for Italian words that have one unique translation for all its senses (cf. column 2 in table 1), whether this translation shares with the Italian word all the CLIPS entries (table 2).

	IT words with same FR equivalent for all their senses	IT words with more than one translation
-aggio	89.9 %	10.1 %
-tà	77.4 %	22.6 %
-zione	80.4 %	19.6 %

Table 1

	FR words sharing the IT CLIPS entries
-aggio	99.97 %
-tà	99.98 %
-zione	99.98 %

Table 2

Table 1 shows that, as predicted, a large number of Italian words ending with the selected cognate suffixes have, for all their senses, one translation ending with the corresponding suffix in French. Table 2 is very striking too and attests that, as predicted, this translation shares with the Italian word all the CLIPS information. The small percentage of errors is due, as expected, to differences concerning the granularity level of both CLIPS and the bilingual dictionary's sense distinction. For example, for *passaggio*, CLIPS has one entry that is specific to the domain of sport (i.e. *a pass of a ball*). In this sense, the word should have a specific translation in French (*passe* instead of *passage*) that is absent from the Robert-Signorelli. In this example, since *passaggio* has, for all its senses, the same translation *passage* in the bilingual database, our algorithm will wrongly infer that *passage* is the translation for this specific sense too. But, as shown by the results in table 2, this situation is very rare. We can then conclude that the cognate approach may allow us to build a very efficient "translation guesser". However, for those constructed words that have more than one translation (cf. column 3 in table 1), our method is inadequate and the *Sense Indicators Approach,* which will be described in the next section, is instead required.

3.2. The Sense Indicators Approach

Good bilingual dictionaries supply TL equivalents for the SL words and help users select the appropriate translation by means of parenthesised words or expressions, or even abridged forms that follow the headword. These sense indicators (henceforth *s.is*) provide in fact syntactico-semantic information that is used as a clue for restricting the sense of the SL item (Atkins & Bouillon, 2003), e.g. *capo* (*promontorio*) → *cap* vs. *capo* (*filo*) → *fil*; *frazione* mat. → *fraction* vs. *frazione* sport. → *relais.* We advocate an approach based on sense indicators claiming that, just as they guide the choice of the adequate translation in a bilingual dictionary, s.is may be used as search keys for identifying, in the CLIPS lexicon, the semantic entry relevant to the Italian sense of an IT-FR pair. We also assert that the semantic similarities holding between translation equivalents – by virtue of their very nature – allow to reasonably envisage that the main semantic properties of the IT sense, encoded in its lexical entry, be eventually shared by the FR corresponding sense.

3.2.1. Experimental Data

The study was conducted on a representative set of 250 nouns and verbs[4] (992 word senses) selected from the CLIPS lexicon population on a frequency basis, privileging highly polysemous lexical units. In the framework of this second approach, the DIF[5], which

[3] Robert & Signorelli : Dizionario francese-italiano italiano francese, Milano : C. Signorelli ; Paris : Le Robert, 2003.

[4] The reason why adjectives were not taken into consideration in the experiment phase is explained at the end of point 3.2.3.

[5] Il Dizionario Francese-Italiano Italiano-Francese, Torino, Paravia-Hachette, 2000.

supplies a great deal of sense indicators and is consultable on Cd-Rom[6], was preferred to the Robert & Signorelli, unfortunately not yet released in electronic format.

3.2.2. Methodology

- □ Extraction from the bilingual dictionary of tuples ITword – s.i. – FRword (henceforth, *X – A – Y*);
- □ Analysis and classification of sense indicators as done by Bouillon and Atkins (2003), according to their nature and frequency of occurrence:
  - Morphosyntactic s.is: verb subclass, auxiliary selection, plural form of nouns, pp type;
  - Inferential s.is: synonym, hypernym, meronym, typical subject or object, domain information;
- □ Distinction between two different types of s.is, according to the way in which the information they provide can be used to identify the CLIPS entry relevant to the IT sense:
  - s.is usable directly, i.e. the string in *A* is searched among the information contained in a CLIPS lexical entry of *X*, e.g.: *X=gioielleria, A=negozio*; *X=fianco, A=lato*;
  - s.is usable upon conversion into the descriptive language of CLIPS, e.g.: if *A* begins with *parte* / *settore* / *ogni*, the corresponding information to be searched for in the CLIPS entries for the lemma *X* is the feature '+part';
- □ Design and implementation of an algorithm whose rules:
  - operate, whenever relevant, the conversion of s.is into the (morphological, syntactic or semantic) descriptive language of CLIPS;
  - retrieve, for each *X-A-Y*, the relevant CLIPS entry of *X*, on the basis of the information provided by *A*;
- □ Controlled assignment to *Y* of the main semantic properties of *X*.

3.2.3. The Algorithm

For deriving a FR entry from an IT one, the first step is, as previously said, identifying the CLIPS entry relevant to the IT sense of the bilingual pair, and whose information we want to ultimately transfer to its FR translation equivalent. To this purpose, the relationship holding between the information content of the entry(ies) of *X* and the clue provided by *A* has to be investigated. Three different types of rules are applied, depending on the nature of *A*:

1) Search for a CLIPS entry of *X* containing the string in *A*[7]:
   - where *A* is, in the CLIPS entry of *X*, the target of a synonymic (A) or a hypernymic[8] relation (B), e.g.: *X=capo*, *A=testa* => retrieved CLIPS entry: USem61397*capo,* whereby *testa* is encoded as a synonym of the headword; *X=capo*, *A=persona…*=> retrieved CLIPS entry: USem3615*capo,* whereby *persona* is encoded as the hypernym of the headword;
   - where *A* is the target of *any* qualia relation (C), e.g.: *X=scuola*, *A=movimento* <=> USem62940*scuola,* whereby *movimento* is target of the relation 'follower_of';
2) Search for a CLIPS entry of *X* sharing properties with a CLIPS entry of *A*. The two entries may share:
   - the target of the hypernymic relation (D) e.g.: *X=comunicare*, *A= notificare*; one entry of *comunicare* and one entry of *notificare* share the target of the 'isa' relation: *dire* => retrieved CLIPS entry: USem6472*comunicare,* whereby *dire* = hypernym;
   - the sortal information (*X* may also belong to either a subtype or supertype of *A*'s type) (E), e.g.: *X= avvertire*, *A=percepire*; one entry of *avvertire* and one entry of *percepire* share the semantic type: 'experience_event' => retrieved CLIPS entry: USem4841*avvertire* (typed 'experience_event');
3) Search for an entry of *X* containing specific information inferred from the conversion of *A*[9] into the descriptive language of CLIPS. Such information may be:
   - a specific semantic type (F), e.g.: *A=gruppo, insieme, complesso* => semantic type (*X*) = Group; Human_Group; *A=persona, chi* => sem.type (*X*) = Human or subtypes[10]; *A=rendere, far* => *X* belongs to semantic types for causative events vs. *A=diventare* => *X* belongs to semantic types for inchoatives, etc.
   - a specific domain (G): *A* ending by a dot is interpreted as a domain information and converted into its corresponding value in CLIPS' domain hierarchy[11];
   - a specific feature or relation (H), e.g.: *A=stare, restare* => Aktionsart (*X*)= State; *A* begins by *per* 'for' or *di* 'of', => search is restricted to the entries of *X* containing respectively a telic or constitutive relation whose target corresponds to the word following the preposition in *A*, e.g.: if *X=asfalto* and *A=per rivestire*, an entry of *X* is searched for whereby the target of a telic relation (here, 'used_for') is *rivestire.*
   - a specific syntactic structure (I), e.g.: if *A*='intr., aus. *avere*, *con avec*', the selected semantic entry of *X* is the one linked to the corresponding syntactic unit encoding a two-place predicate with a pp_*with* complement; *A*=pron. (*se stesso*) => selection of a reflexive structure; *A*=pron. (*reciprocamente*) => selection of a reciprocal structure, etc.

The matching rules (A) to (I) were ranked from 1 to 9 (cf. table 3) according to the degree of reliability the information types they are based on confer to the results, and applied in such order. The rule application order is in fact crucial to the correctness of the algorithm: the higher the rule rank, the more reliable the result. The probability that the retrieved entry of *X* be the correct one is higher if the query is based on the information concerning its domain of use or its syntactic structure (rules G or I) than if it is grounded on the properties this entry shares with an entry of *A* (rules D or E). In fact, such shared properties (hypernym or semantic type) might be far too generic to be significant. Likewise, rule (C) might lead to misleading results in case the relationship holding between *X* and *A* is too weak.

Of the whole set of bilingual pairs under study, 96.16% had a sense indicator in the dictionary and thus could be accounted

[6] a Cdrom-based extraction of data is obviously to be regarded only as a provisional solution, for the feasibility study phase: more rapid and straightforward extraction facilities are needed that require the publisher's involvement.

[7] *A* may consist of a single word, two words separated by a comma or a succession of strings. In the latter case, the first string is taken to be a hypernym.

[8] synonyms of hypernyms and hypernyms of hypernyms are also taken into account.

[9] or the first string of *A*, in case *A* is a succession of strings.

[10] or X contains one of the semantic features: +part, +collective, +human

[11] the same holds for s.is expressing domain information with a different wording, e.g. *nel calcio*.

for by the algorithm. Applying the above rules to this subset, we obtained a recall rate of 64.9% (IT sense of a bilingual pair successfully linked to the relevant CLIPS entry). As table 3 shows, it is through the very application of the higher level rules that the higher percentages of good results were achieved.

rule type 1			rule type2			rule type 3		
A-1	B-2	C-9	D-7	E-8	F-6	G-3	H-5	I-4
16.6%	26.8%	0.92%	8.9%	5.8%	3.9%	12.3%	9.2%	15.4%

Table 3 Distribution of success rates over the rules

Besides these successful results, another 4.16% of links were obtained by means of a default rule linking a unique sense to a unique entry. Moreover, out of the 30.94% of failures, in 7.74% of the cases two theoretically possible results were returned, which could undergo manual disambiguation.

The success of rule application clearly presupposes a coincidence between the granularity level of both CLIPS and the bilingual dictionary's sense distinction. Wrt. our task's purpose, the problems encountered with the DIF were essentially due to an excessive splitting of senses, to a prevalence of collocators (unexploitable for the matching against CLIPS information) over synonyms to discriminate adjective senses[12], e.g.: 1. *acuto (di suono)*; 2. *acuto (vista)*, and to some inconsistent information marking, i.e. same information expressed by different wordings throughout the lexicon. By contrast, cases of failed matches due to inexistent IT senses or to a wrong assignment of semantic properties were imputable to CLIPS encoding.

In the following section, the combination of the two methods will be evaluated.

4. Global Assessment

Comparing the two methods, it clearly appears that the cognate-based one is easier to apply and yields a higher recall rate as to the number of CLIPS entries that are linked to a French word. It is, however, far less precise than the method based on sense indicators, since the only inference that can be done is that the FR word with the corresponding cognate suffix will share all the CLIPS entries of the Italian word, e.g.: *rigidité* shares all CLIPS senses of *rigidità*. Should one sense in the dictionary have synonyms, e.g.: *rigidità*, (1) *rigidité*, *dureté*, (2) *rigidité*, *raideur*, these would not be linked to their CLIPS entry. By contrast, the s.i.-based method is far more demanding and yields a lower recall rate but it does enable to relate various different FR equivalents to a given IT word or CLIPS entry. A priori, all the translations equivalents could potentially be linked. Table 4 below sums up the number of CLIPS entries that can be linked to at least one FR equivalent by combining the two methods for *–tà*, *–zione*, and *–aggio* suffixed nouns:

handled by cognate approach		handled by s.is approach	
correct	failed	correct	failed
82.54%	0.02%	11.71 %	5.73 %

Table 4. Combining both approaches for handling constructed words

These results call for a few comments. On the one hand, the complementarity of the two methods is striking; of the17.46% of words to which the cognate method does not apply, 11.71% are successfully handled by the s.is method. The success rate for the handling of the suffixed words under study is thus 95%. On the other hand, attention should be drawn on the reliability of the cognate-based approach. If we consider, as Dubois (1971) that constructed words represent 68.2% of the vocabulary and could therefore be potentially handled by the two combined methods, these results become very encouraging. As to non-constructed words, the total success rate (matching and default rule) is 69% and since the algorithm is strongly dependent on the information provided by the bilingual dictionary, this rate could be further increased by gleaning the most informative data from different sources.

5. Concluding Remarks

Deriving new lexical resources from existing ones is undoubtedly a worthwhile venture both in terms of time and effort. The building process is in fact simplified and shortened, benefiting from the achievements of previous research while obeying the undisputed principle of reusability of existing resources. The input lexicon also profits from such a practice that implicitly entails the assessment of both its coverage and coding consistency.

The study presented in this article is meant to lay the foundations for a more extended research. The experiment performed has yielded promising preliminary results that encourage us to carry on, even more if we consider that the two approaches taken are applicable to other language pairs sharing similarities in terms of morphological structure. Italian and French monolingual lexica could then constitute the basis for the development of a bilingual lexicon.

References

Atkins S. and Bouillon P. (2003), *Relevance in Dictionary-Making*, in Proceedings of ``I simposi International de Lexicografia", Publicacions de l'Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

Geysen, R. (1990), *Dictionnaire des formes analogues en 7 langues avec résumé de grammaire comparée*. Paris, Duculot.

Lenci, A., *et al.* (2000), *SIMPLE Linguistic Specifications*, Deliverable D2.1, ILC-CNR, Pisa.

Namer, F. (2001), *Génération automatique de néologismes bilingues morphologiquement construits en français et en italien*, TALN 2001, (pp. 281--296), Tours.

Pustejovsky J., Boguraev B. (1993), *Lexical Knowledge Representation and Natural Language Processing*, Artificial Intelligence 63, 193--223.

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

Pustejovsky J. (1998), *Specification of a Top Concept Lattice*, ms., Brandeis University.

Ruimy N., *et al.* (1998), *The European LE-PAROLE project: The Italian Syntactic Lexicon* - First International Conference on Language Resources and Evaluation Proceedings, vol. I, (pp. 241--248), Granada.

Ruimy N., *et al.* (2002), *CLIPS, a Multi-level Italian Computational Lexicon,* Third International Conference on Language Resources and Evaluation Proceedings, Vol. III,(pp.792--799), Las Palmas de Gran Canaria.

[12] The inadequacy of the s.is for adjectives does not bear on the feasibility of this approach: the use of a different dictionary, e.g. the Robert & Signorelli, would in fact allow to overcome the problem.