French WordNet progress, And Structured concepts embodiment inside Wordnet.

Dominique Dutoit
MEMODATA
LITIS (Univ. of Rouen)
CRISCO (Uni. of Caen)
do.dutoit@gmail.com

Patrick de Torcy
MEMODATA
17 rue Dumont d'Urville
14000 CAEN
p.detorcy@memodata.com

Yann Picand
MEMODATA
17 rue Dumont d'Urville
14000 CAEN
y.picand@memodata.com

Abstract

In this paper, first, we will give some global measures of the French WordNet progress and, second, we suggest the introduction of new relations inside WordNet which are part-whole relations. In that part, we study the benefits of these relations in morphology, syntax and semantics, including some intensional semantics.

1 Introduction

This paper deals with the embodiment of structured concepts, such as part of grammar or meaning inferences, inside the wordnet structure.

As many researches, this particular research is based on the substrate outputs of the past works. Working since 1989 in the field of lexical semantics, of course this substrate contains different layers which apparently could use specific algorithms to be accessed and played. One goal of this embodiment is to unify all these algorithms in a single one, allowing us to converse between distinct steps of the analyses. Focusing on the structured concepts introduction techniques inside wordnet, we will not deal about its necessity. Nevertheless, we will have to remember the main algorithm and some particularisms of our ongoing concepts and relations. The given work is based on *The Integral Dictionary* (TID, [Dutoit 1992]), a French semantic net, which is today fully merged in Princeton' WordNet. The older infrastructure of TID is based on the structuralistical point of view called *componential semantics.*

Then, the paper plan is:

- Data input : sample concepts
  - The French Wordnet inside TID
  - Other models
  - Algorithm
- New : structured concepts
  - Formalism
  - Some use cases

This proposal about the introduction of structured concepts inside a graph was announced and argued in a long academic document (in French) [Dutoit 2009] which is a "Habilitation à Diriger des Recherches" (Habilitationschrift).

2 French data input : sample concepts

We were a partner for the French in EuroWordnet [Vossen 1999], and in Balkanet [Stamou 2002]. In this section, we summarize the current state of the French WordNet with quantitative measures, we give details about Wordnet and TID integration, we give details about specific TID formalism and we give a short view of the algorithm used in text analysis

2.1 Current state of the French WordNet

Table 1 shows the main figures:

Princeton synset with a French Literal	79.562
Princeton synset with a French gloss	38.252
Princeton synset with a French label that is not a literal (an entry for the French dictionary) but only an explanation or a suggested translation	2.357

Table 1: the French regular Wordnet

2.2 WordNet and TID integration

Originally (before 2000), TID had not the benefits of the synset object. Then, once introduced this structure, TID was able to share its data with WordNet. Having the synset compounds, the

notion of gloss was also reachable. Fortunately, we had an explanatory dictionary for French not linked. Then, we use some heuristics (see below the main one: semantic distance), to fill semi-automatically the French Synsets derived from Princeton. This approach yields to the statistics given in Table 1 relating to the French glosses.

Today, TID is fully compatible with Wordnet. But, as TID contained some other layers, of course, we maintained these data. The other layers was derived from the Meaning-Text Theory (MTT) (Melcuk, 1996) which is influenced by a generative point of views and from the componential theories (for instance Pottier, 1992) that are pure structuralism.

2.3 The MTT model inside TID

Here, we will not explain the whole model but only some features related to the aim of this paper.

The MTT model manages some formal RELATION between word-meanings. A word meanings is a kind of word, but not a kind of synset. In TID, we decided to allow MTT relations between: word-meaning and word meaning, word meaning and synset, and synset to synset, due to some needs in terms of redundancy and accurateness.

All the MTT relations are more precise than the wordnet relations between synsets. For instance, where Wordnet defines a bidirectional relation between one *improvement synset* and one *improve synset*, the MTT model defines a directional relation between them (in this example, the nominalization case with a formal label meaning *action of* or *result of*).

Most of the MTT relation are not included into the wordnet model or TID, and reify some complex relationship between terms, hiding inside a partially composite formal label a kind of structured complex concept which is one aim of this paper. Then, the MTT relations are formally only binary relations, and can not define directly structured forms.

Table 2 shows the main figures concerning the implementation of MTT in TID (and by inheritance in the French WordNet):

Number of different MTT types of relations	45
Number of MMT relations between synset or word meanings	112540

Table 2: the French Wordnet implementing MTT

Of course, when the French synsets are merged to the Princeton synsets, Princeton synsets inherit of the improvements given by MTT.

2.4 Componential models inside TID

All the componential models create a formal distinction between words or forms and meanings or formal concepts. Then, TID which is an empiric work originally induced by this saussurian distinction, contains some objects that are not words but only "concepts". This paper has not the goal of dealing with advantages or disadvantages of this dualism. Sometimes the formalism seems heavy, sometimes it is clear that it allows the description of some particular features, for instance when "natural" hypernyms are not available in the lexicon of a given language.

Then, with that distinction, we have to accept that could exist:

- a word meaning "horse-equine",
- a synset "horse-equine",
- a class of "horse-equine",
- a kind of topic "horse-equine"
- and some other things like this.

Figure 1 gives some details about this:

Relation	**Node**
-	cheval-horse (synset) (1)
generic (2)	horse (class) (2)
spec, encyclop. (2)	animal breeding (class) (2)
spec, encyclop. (2)	mount (class) (2)
spec, encyclop. (2)	beast of burden (class) (2)
spec, taxonomic. (2)	equid (familly of) (class) (2)
To topic (2)	Horse (theme) (2)
holo_member (3)	type genus of the Equidae… (synset) (1)
hypernym (4)	hoofed mammals … (synset) (1)
derivative (3)	provide with a horse … (1)

Figure 1: Wordnet, TID and MMT sharing their details

Legend:

1: WordNet and TID
2: only TID
3: only WordNet
4: WordNet and MMT

Have a look to http://dictionary.sensagent.com/cheval/fr-en/#analogical to get a better view of these details. Let's notice that on this site, most types of relation have been simplified and/or merged.

Table 3 shows the main figures relating to the implementation of a part of componential semantics in TID (and by inheritance in the French WordNet):

Number of different TID concepts	23
Number of TID concepts	40.000
Number of different TID relations	36
Number of TID relations	340.000
Number of WordNet Synsets linked to TID	see *table 1*
Number of TID synsets for French	220.000

Table 4: Componential semantics in TID

2.5 One Algorithm

The algorithm goal is to reach a subset of concepts (or sometimes synsets) which links two words-meanings. For instance,

samurai & warrior	*Have to select*	warrior (class) warrior (synset)
samurai & saber		war (theme) fight (theme)
samurai & katana		war (theme) fight (theme) Japan (theme)
samurai & Tokyo		Japan (theme)
samurai & to ennoble		nobility (theme)
florist & tulip		flower (theme)
florist & to sell		trade (theme)
florist & to exchange		exchange (theme)
florist & person		person (class) person (synset)
florist & shop		shop (theme)

Figure 2: Expected concepts extracted from TID and/or WordNet

In figure 2, we can observe that the extracted concepts are not relations between the two terms but a (A) sample abstraction (common compound, componential compound, semantic feature, analogy) which subsumes the two given words(-meaning).

It is important to notice that most of the extraction could not be found in a concrete text inside a local area, such as a phrase or a short sentence. The good unit to discover these terms together is often a large portion of a discourse. For instance, a sentence like "florist sells flowers" is uncommon, except, of course, in a dictionary. Then, (B) a dictionary could not explain easily: "the advocate sells his pen".

In this condition, the integration of structured concepts could have the both A and B roles.

Finally, we have to give an idea about the algorithm itself. We proposed an original way to measure the semantic proximity between two word senses. This measure takes into account the similarity between words (their common features) but also their differences. It was described first in [Dutoit 2000]. Also this article could give some details about some uses in concrete text [Dutoit 2002].

Comparison between two words is based on the structure of the graph: the algorithm calculates a score taking into account the common ancestors but also the different ones. The notion of "nearest common ancestor" is classical in graph theory. We have extended this notion to distinguish between "symmetric nearest common ancestor" (direct common ancestor for both nodes) and "asymmetric nearest common ancestor" (common ancestor, indirect at least for one node).

Definition: Distance between two nodes in a graph

We note d the distance between two nodes A and B in a graph. This distance is equivalent to the number of arcs between two nodes A and B. We have `d(A, B) = d(B,A)`.

Let's say :

h(f) = the set of ancestors of f .

c(f) = the set of arcs between a node f and the graph's root from the point of view of f.

Definition: Nearest common ancestors (NCA)

The nearest common ancestors between two words `A` and `B` are the set of nodes that are daughters of `c(A) ∩ c(B)` and that are not ancestors in `c(A) ∩ c(B)`.

It is possible to define a measure to calculate the similarity between two words from these sets. We call this measure *activation* (see Dutoit 2002), but as this paper is not focused on this measure we will not give more details. Let's remember that it's possible to use the activation to measure the semantic proximity between two word senses following a particular *point of view*.

As this paper is focused to introduce some other points of view on the net (then: other type of concepts, not only shared semantic feature), we work on the hypothesis that *the same algorithm could be used to extract these new concepts in a larger, more heterogeneous set of NCA*.

Another interesting subset of shared concepts could participate to the measure of the smaller differences between meanings.
For instance, the smaller differences between *samurai & Tokyo* have to be*:*
d(samurai, Tokyo)=f(aristocracy, warrior)
d(Tokyo, samurai)=f(town)
whereas similarity is immediately a function of *Japan*.
This anti-symmetric measure of differences is particularly interesting for several tasks, such as calculation of the saturation of a word or a set of words by another word or another set or words. It used the subset of *Asymmetric nearest common ancestor (ANCA).* The asymmetric nearest common ancestors from a node `A` to a node `B` `are` contained into the set of ancestors of `c(B) ∩ c(A)` which have a direct node belonging to `h(A)` but not to `h(B)`.

2.6 Some uses of the algorithm

The algorithm was used to:

- measure one semantic weight of each word in a text related to the other words of the text
- make a reverse dictionary service computed from the graph (TID, WordNet etc)
- compare sentences, for instance, to merge TID with WordNet

We think that if we introduce in the net some other concepts, which will be structured concepts, then several other services will be reachable.

3 Introducing structured concepts and dynamic concepts

This section defines structured concepts and some rules to reach and saturate them. The most effective use of these structured concepts, when they are validated, is to generate some dynamic concepts which were not directly reachable before their activation. A deep discussion about these concepts, in philosophic, linguistic or computational terms can be found on Dutoit [2009].
To discuss these notions, we will introduce the following simplest examples:

- compound words
- finite expression
- infinite expression

and, more essential for our conference:

- lexical semantics.

3.1 Compound words

The French compound *pomme de terre* is a translation to the simple word *potato.* Of course, it is possible to design a particular compound term processor to localize this expression in a given text. But, this tool can not decide itself, with its privative morphological knowledge if the right interpretation of the compound is the compound itself (the whole) or its parts. For instance, is a *lung neoplasm* a simple term or a compound term? It is related to the point of view. Interpretation of *pomme de terre* is the same. For instance, in term of lexical reduction, you get [*pomme de terre*] in French and can get [*potato*] in English. But in term of grammatical rules you have to consider 3 separate words. Introducing structured concepts in our graph (TID, WordNet) could answer to this instability of these points of view using an instable graph containing several structured concepts.

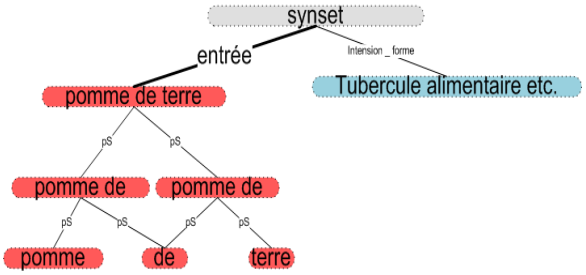

Figure 3: a compound term

This graph defines a partially ordered set which is a lattice. The binary character of the graph is influenced by the algorithm presented above. Using this algorithm, it appears that the entry point to the analyzers becomes the lexicon itself and not the syntax. In that approach, syntactical constraints (such as relative token position, belonging to a given chunk or able to support a particular syntactic relationship) are applied when the (structured) concept is detected by the lexicon, and not before, as a result of these rules. Then, the algorithm stays unchanged, but, in the cases of structured concepts, augmented by a constraint section.
Let's highlight the procedure.

Structured concept

The part of the figure 3 with relations labeled e_i, where e means *element* and *i* is an index.

Dynamic concept

The part of figure 3 with a relation labeled *entrée*, where entrée means that the created whole *pomme de terre* is also an entry of the dictionary.

Formally, we do not use the relation label: *entrée*, but the label λ_{t_whole}, where *t* means "token" and *whole* means that the whole *pomme de terre* could be considered as a simple *token* **in** the point of view found by the structured concept.

Remark: *the area of the structured concept presented above is very close to the popular Google index.*

3.2 Finite expression

Let's take the example of a date.

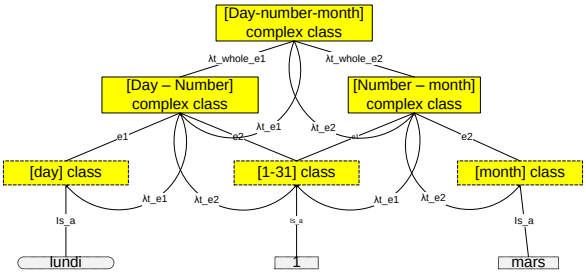

Figure 4: building a date using the classes of the graph

This graph does not introduce a strong innovation except lambda calculus. The previous one was simpler due to absence of classes. With presence of classes, of course, the built date is not (in the French order) [[*day] [number] [month]]* but, in this example [[lundi] [1] [mars]] (literally *Monday 1 March*).

Structured concept

In the figure, [day-Number] is a structured concept (it needs one day name and one particular number, with a specific syntactical constraint) to be activated. The figure shows 3 structured concepts like [day-Number].

Dynamic concept

In that figure, when satisfied, [day-Number] emits the order of placing 3 identified tokens:

- λt_e1 : *lundi*, as a token, which means that *lundi* in *lundi 1* is a day. This result is not innocent. For instance, *lundi*, in a phrase like *the noun lundi*, has to be *viewed* as a noun and not as a (for instance : hyponym of a) day.
- λt_e2 : *1*, as a token, which means that *1* in *lundi 1* is a number of day. This result is not innocent. For instance, it should very useful to "understand" a very short dialog like:

Laura: Could you come "lundi 1"?
Elie: No, "le 1",I will be very busy.

In fact, this process creates a local extension of the date concept that is not admissible in a dictionary because *1 may co-refer everything* except when you have spoke about the day *lundi 1 Mars* or the horse number 1, or the boat number 1 etc.

- Λt_whole_e1 : *lundi 1*, as a token, which means that *lundi 1* in *lundi 1 mars* could be considered as a non-dependant thing and could be re-used as is.

3.3 Infinite expression

The case is given by the evaluation of arithmetic expressions which are not aprioristically sizable. Dutoit 2009 shows how to close the expression in the graph before attempting the calculation.

3.4 Several phrases

We are going to deal with the following phrases:
a) the noun samurai a') the noun fortunately
b) a white horse b') a white thing
c) the color of that horse c') the color of that thing

Cases (a) and (a')

Case (a) interest is that it exists somewhere a place where this particular phrase has a meaning. Unfortunately, this place which establishes links between POS and words is not reachable for the NCA: assuredly, the point of view of componential semantics (based on shared part or meaning) is different to the POS point of view. But, this difference does not modify deeply the relations nature. If we say:
samurai has one hypernym *warrior (generative point of view)*
samurai is a kind of the concept [*warrior*] (componential point of view)
samurai is a kind of [*noun*] (POS point of view), we do not say the same thing at all.
But, if we examine the case (a'), it appears that the previous discourse is not sufficient. For instance, the sentence *the noun fortunately does not exist in English* is trivial. Then, what could be the nature of the relationship between *noun* and *samurai* or *fortunately*? To simplify our answer, we can remember a very classical thesis due to Aristotle. For our example, his analysis should be:

- in (a) and (a'), *noun* is an active power and predicated terms are passive powers.
- in all situation, one active power tries to apply its program to its environment, and success of this procedure is in dependence to the aptitude of the passive power to support the active power transformation.

In our example, *samurai* which have an active power in other situation, here, is only a passive power that can support the transformation (point of view, a focus to one part of its whole) *noun.* Of course, *fortunately* does not support this in our current state of knowledge in English.
Then, the mechanism of this structured concept is something like:

noun previous (a_word) → try to consider a_word as a kind of noun.

Then, one **v**irtual instance of **a_w**ord (**va**_word) is created has a kind of *noun.* If both elements a_word & vaword exist together in that area, the related NCA will be returned.
Figure 6 shows this situation:

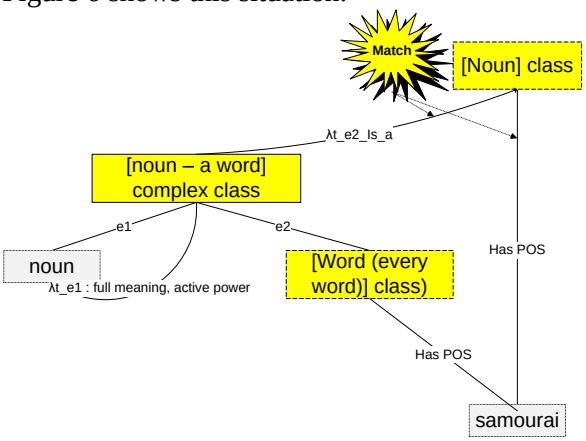

Figure 5: building the intension of *the noun XXX* using the classes of the graph

Structured concept

In the figure, [noun- a_word] is the single structured concept and it reifies one active power of meaning the word *noun.*

Dynamic concept

Elsewhere, [noun] class is both a static concept (because it returns Noun as POS when accessed by a potential noun token via the dictionary entry *samurai*) and a dynamic concept because it could be filled by a session token which is *samurai: noun* in the phrase the *noun samurai.* It is a first very short example of intensional semantics (not intentional in the Husserl use in phenomenoly). Von Fintel [2002] write in that purpose:
We need to move to a semantics that is intensional in the following sense: it has to contain operators, like former, that "displace" the evaluation of their complements from the actual here and now to other points of reference …

Case (b)

The most frequent approach that deals with these phrases consists to assume simply world knowledge where a horse has a color property or can be colored. As we have read it, these considerations are not based on lexical semantics (for instance, explanatory dictionary) but on our world experiment. Then, this approach is close to CYC (Lenat, 1999) and far from WordNet. Is there somewhere in WordNet a source that could compute a *linguistic* relation between *color* and *horse*, or *white* and *thing?* We suppose that this source is inside the gloss itself, and could be computed if we improve the formal representation of the glosses.
Reading WordNet gloss, it is visible that *color* is a visual appearance of an object. To simplify the discourse, we consider that this object has a strong opacity, and, then, we suggest the part of an object that gives this visual appearance is its surface part. Then, we can try to define again this meaning of *color*. It could be*: a property of the surface of an object that...*
At this point, it appears that *color*, similarly to the word *noun*, in certain position, has an active behavior that moves a *referent* word, like *horse*, from its places (given by its definition*)* to another place like a *surface*, or more precisely, like a *thing viewed from its color.*
In figure 6, the structured concept [color of a_noun] emits the fact that the **token** *horse* **could be viewed** as a [thing viewed for its color]. By inheritance, such thing **could** reach, **could** be activated as a [surface], a [volume] ... and finally as a [concrete thing]. In another point of view, a general linguistic knowledge inherits from the fact that an animal could have a body (body is defined in WordNet as *a entire structure of an organism*) that the **original** token *horse* could be viewed as an [equid] ... an [animal] and finally as a [thing having a body] (in the definite meaning).
At bottom of figure 6, `λt_property(e2,e1):Is_a` (1), where (1) is a comment, creates a new token *property*(*horse,color)* with this place.
At top of the figure, you have the expression `λt_whole.` This expression is a concatenation of the new token *horse* and the older token *horse:color*. It is a kind of identification. Then,

- `λt_whole:Is_a` (2) claims that *horse_ property*(*horse,color)* is an attested member of [concrete thing]
- `λt_e2:Is_a` (3) claims that *horse_ property*(*horse,color)* is also an attested member of [thing having a body]

Rem. 1: in our example, a such *horse_ property*(*horse,color)* selects the different meanings of *horse* that are defined as an animal,

but a similar expression will select some other meanings for *horse*, such as *heroin*, in another similar path.

Rem. 2: the stranger expression horse_ *property(horse,color)* could be considered as a mean to memorize that these understandings of *horse (concrete thing* and *thing having a body)* have been **simultaneously** checked by two paths, where the first one is the predication (which is not enough) and the second one is a linguistic inference resembling to a world knowledge (which is not enough). Before this simultaneous event, nothing exits (no latent ambiguity), after this simultaneous event a new thing exists with two congruent interpretations available for future calculations.

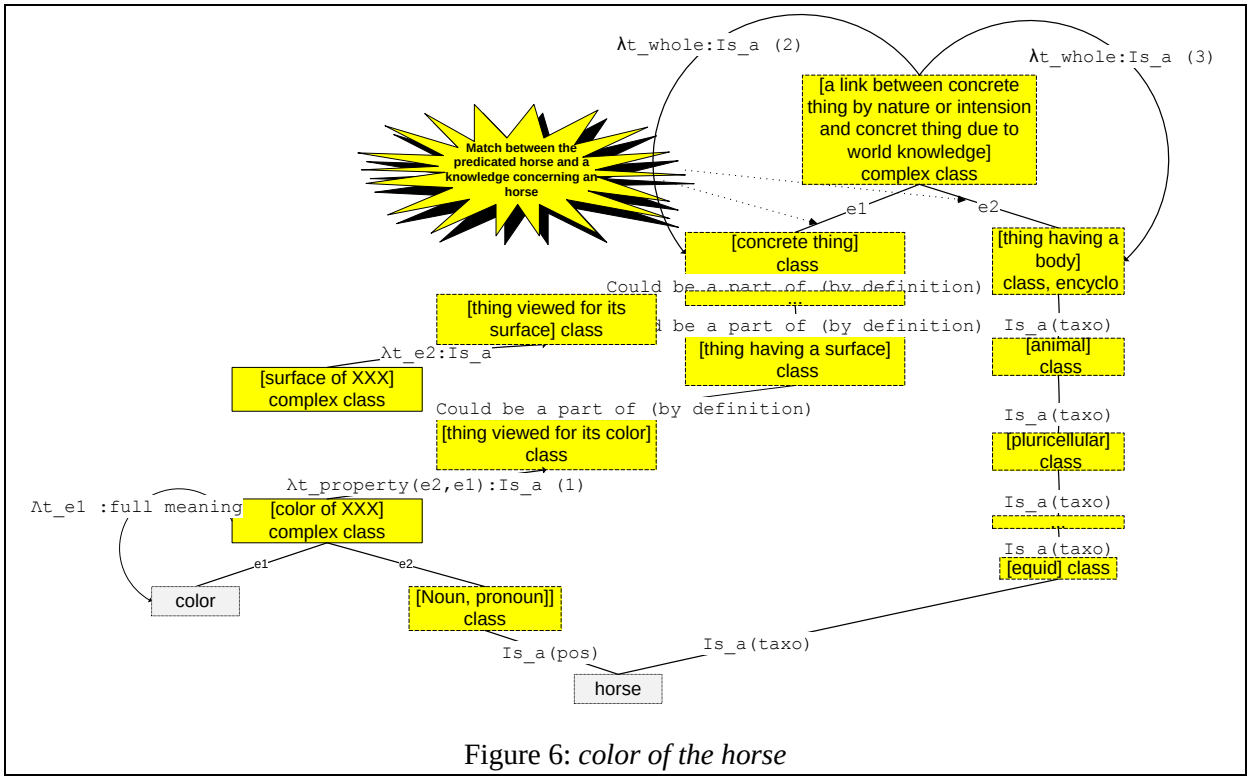

Figure 6: *color of the horse*

It is not possible to draw in a single figure, the dictionary and its running effects. The built instances will be shown at the end of this section.

Case (c)

Drawing *white horse* is very similar with *horse color* in the direction of concrete thing. Nevertheless, we have a particular task with this phrase. As our goal is to embody a part of the meaning (the glosses) inside the graph itself, we have to implement the fact that *white* invokes itself the concept *color* as top attribute. Bottom of figure 7 shows the relation: *e2, added*. Role of this relation is to allow such invocation.

The view of instances

Lambda operators have the role of moving tokens but also of creating tokens. It could be useful to draw this result. Figure 8 shows this result. As mereotopology [Smith B., 1996] is one source of this work, the resulting figure is not a surprise: the part-whole relations are fully maintained.

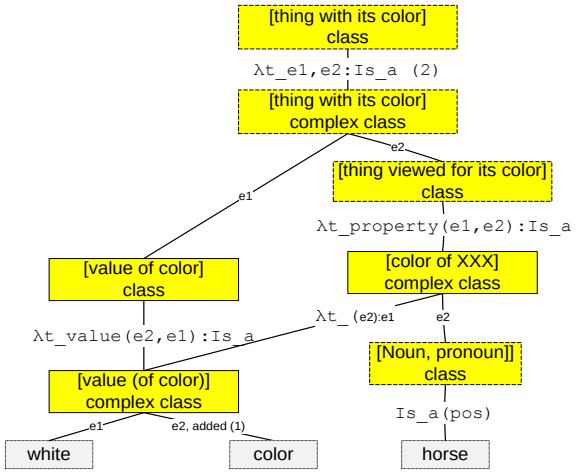

Figure 7: *white horse*.

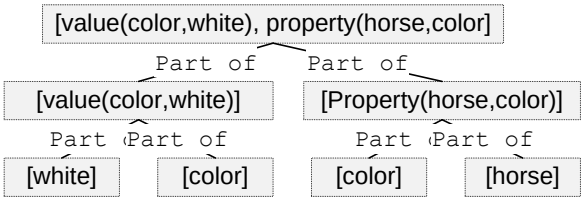

Figure 8: instance for *white horse* **.**

Case (b') and (c')

Could a thing support *color* or *white* as predication? If we consider *thing* as *anything*, the answer is *no*, considering that only particular things (such as *concrete thing*) could support the predicate. Therefore, it is why *color of the thing* could not, in our opinion, be directly computed. As we have to localize the [concrete thing] subset, it appears that an old mechanism could help us in this matter: it is the reverse dictionary technique [Dutoit, 200]. Contrarily to the previous work that runs from bottom to top of the graph, the reverse dictionary runs partially from top to bottom in order to induce *concrete thing* subset from the things set.

A funny application of the built graph

Words after words, the graph builds more complex instances linked together by mereotopological relationships, as in *pomme de terre*. We have shown [Dutoit 2009] that playing the graph on a question like "what is the color of a white thing?", after a formal introduction or one meaning of *what*, could finally creates the following tokens:

```
[
 [value(color,white), property(thing,color)]
 [value(color, ?what ), property(thing, color]
]
```

where *?what* and *white* co-refer a same thing.
This result is due to the fact that it does not depend to external knowledge (world knowledge, local knowledge or reverse dictionary), but only to a lexical semantics formalization of the explanatory dictionary glosses.

4 Conclusion

In a first part, we have described the French WordNet current state in various statistic tables. In that part, we have also summarized the environment of the French WordNet, which is strongly linked to other works in lexical semantics, where all these works make together the body of *The Integral Dictionary (integral* means this integration of theoretical point of views*)*. In a second part, we have suggested improvements of WordNet formalism by introducing some new relationships which are part-whole input / output relations. We studied the opportunities of these relations with various lexicon problems, from morphological area to naive world knowledge area. Introduction of these relations was both motivated by improving the static model of the dictionary, including an *intensional* point of view which is then available for a dynamic building of phrases "understanding".

Reference

Christian Fellbaum, 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT press.

Dominique Dutoit, 1992. *A set theoritic approach to lexical semantics,* COLING.

Dominique Dutoit, 2000, *Quelques opérations sens→texte et texte→sens→texte utilisant une sémantique universaliste apriorique,* PhD, Greyc, Univ. of Caen.

Dominique Dutoit, Thierry Poibeau., 2002. *Combining knowledge sources for resource acquisition*. (COLING), Taipei.

Dominique Dutoit, Pierre Nugues, 2002. *The right word*, full paper, LREC, Las Palmas

Dominique Dutoit, 2009. *Intégration structurale des points de vue componentiel et compositionnels*. HDR, Univ. of Rouen *:* http://www.memodata.com/2004/fr/publications.shtml

von Fintel, 2002. *Lecture Notes on Intensional Semantics,* http://www.phil-fak.uni-duesseldorf.de/summerschool2002/fintel.pdf

Douglas B. Lenat, 1999. *From 2001 to 2001: Common Sense and the Mind of HAL*, http://www.cyc.com/halslegacy.html.

Igor Mel'čuk, I. 1996. *Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon*. In L. Wanner).

Bernard Pottier, 1992. *Théorie et analyse en linguistique*, Coll. Hachette Supérieur.

James Pustejovsky, 1995. *The generative lexicon*, Cambridge, Mass. MIT press

Barry Smith, 1996. *Meretopolopy: a theory of parts and bundaries,* http://ontology.buffalo.edu/smith/articles/Mereotopology1.pdf

Piek Vossen, 1999, *Final report*, EuroWordNet, LE2-4003, LE4-8328.