MWEs in the FRMG framework

Éric de la Clergerie
<Eric.De_La_Clergerie@inria.fr>

http://alpage.inria.fr
INRIA Paris-Rocquencourt / Univ. Paris Diderot

AIM-WEST & PARSEME-FR Workshop
Grenoble, 3–4 October 2016

FRMG for dummies

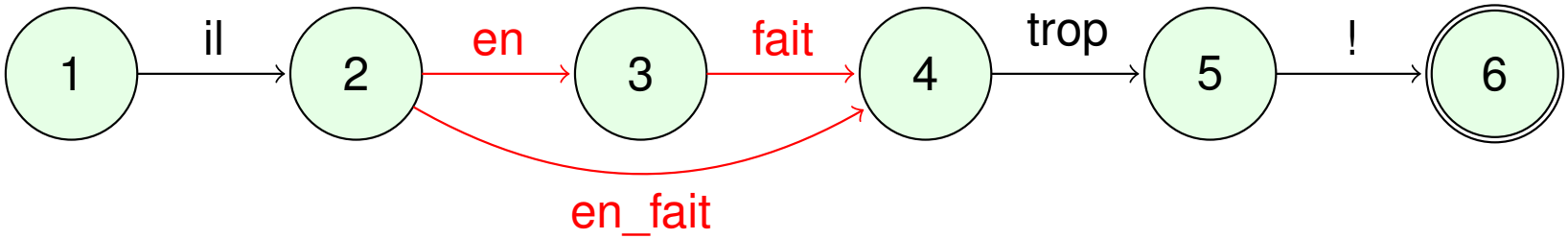
Since 2004, FRMG, a large coverage TAG for French

- generated from a meta-grammar
  - with elementary trees built from constraints
  - and constraints from classes, that inherit from ancestor classes and may be combined
  - $\rightsquigarrow$ compact grammar thanks to tree factorization (381 trees)
- as input, a word lattice (DAG) built by SxPipe
  $\rightsquigarrow$ keep lexical and segmentation **ambiguities**
- as output, **all** (full or partial) parses as share forests
  (derivation forest then dependency forest)
- Disambiguation phase to get a dependency tree, using
  - hand-crafted rules with hand-crafted weights
  - now weight tuning by learning from French TreeBank (FTB)
    + attachment affinities learned on large corpora (distributional hyp.)
- $\sim$ 97% full parse coverage on FTB, and $\sim$ 88% accuracy (LAS) on FTB test
- try it on FRMG wiki at `http://alpage.inria.fr/frmgwiki`

Some « easy » MWEs

Most MWEs are actually handled by SxPipe,
as lexical entries found in Lefff lexicon (adv, det, prep, csu, nc, . . .)

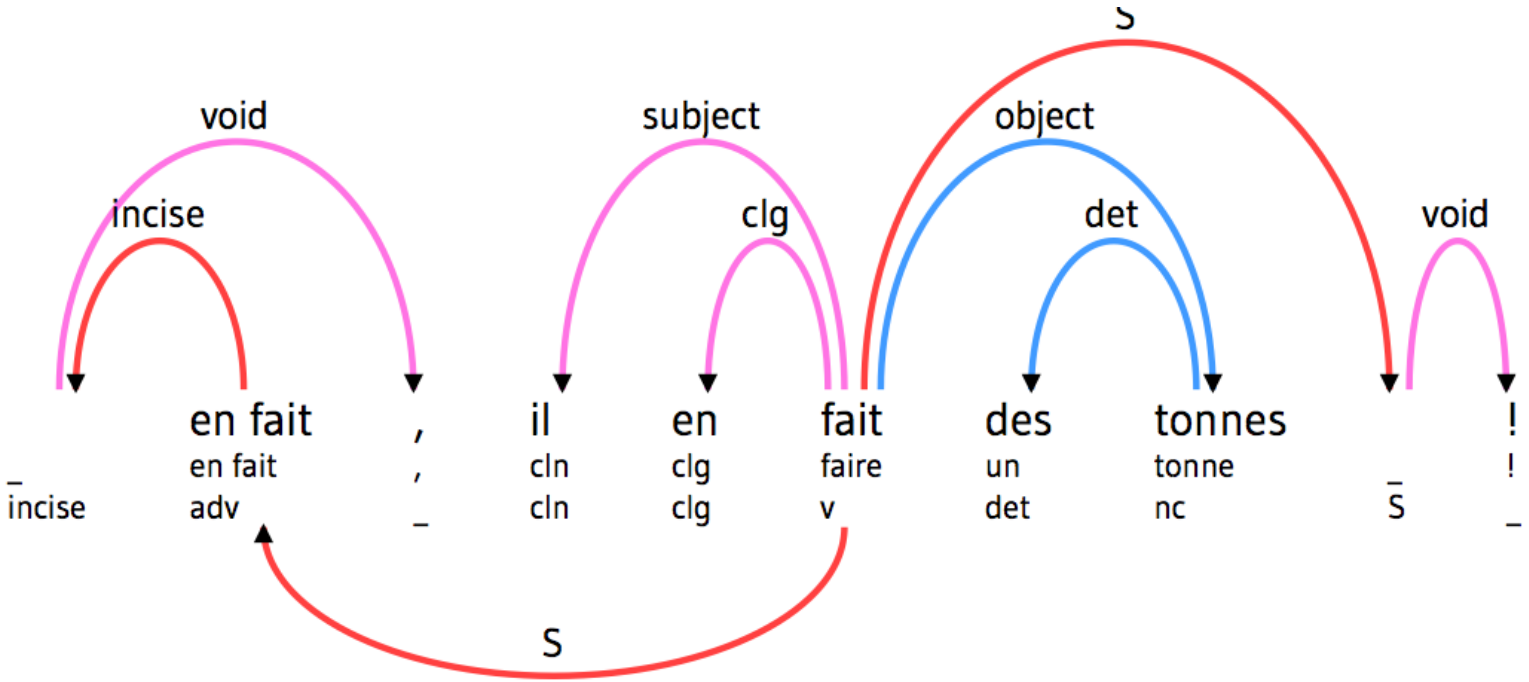
Generally, we keep an ambiguous reading in word lattices

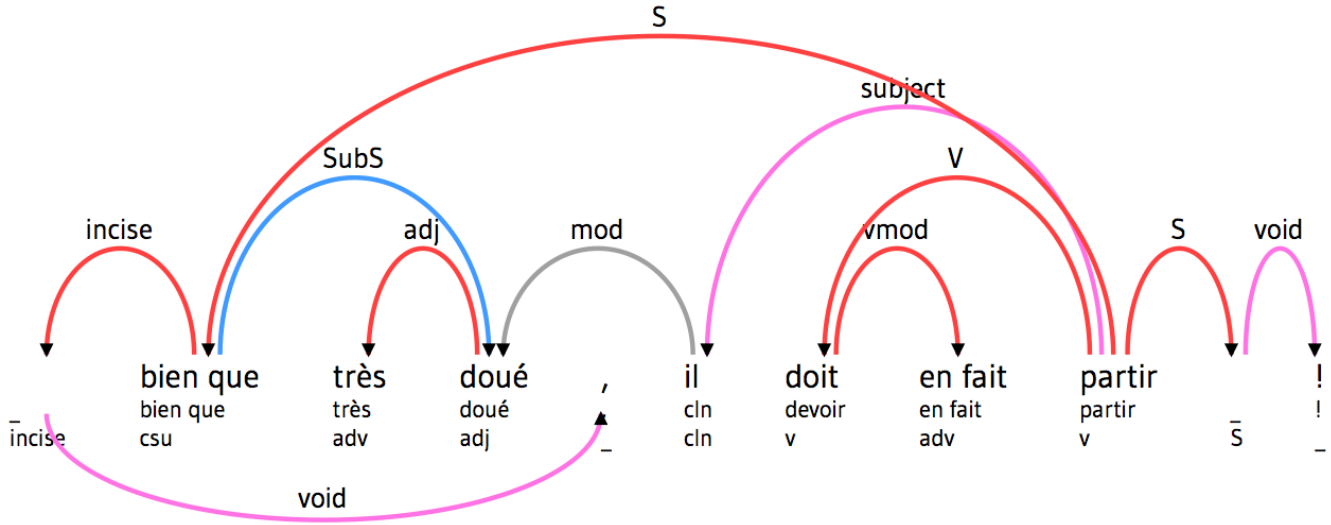


Parsing selects the valid readings, with or without MWEs

By default, the post-parsing disambiguation phase favors MWEs
(less clear after tuning)

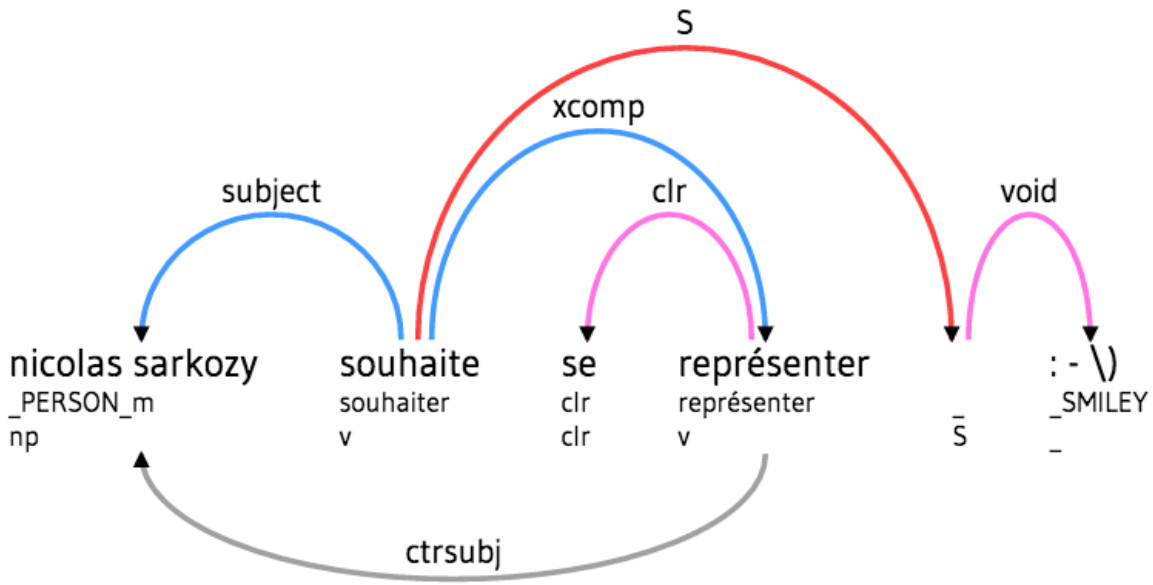
Easy MWEs (cont'd)





Named Entities

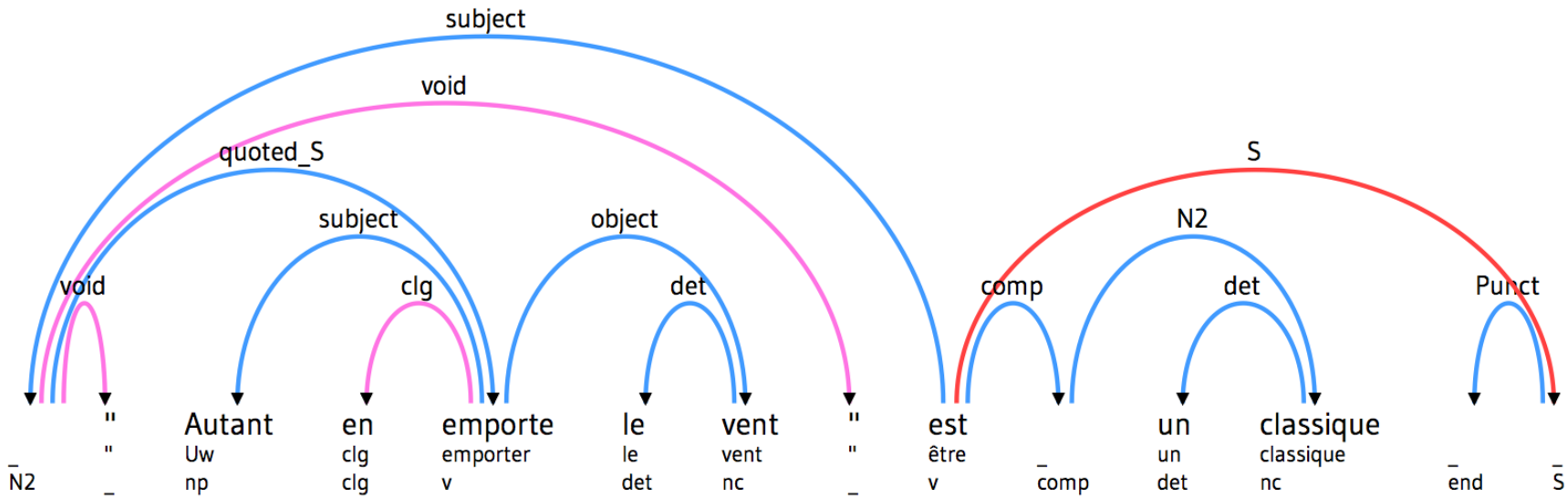
also provided by SxPipe : similar to the previous scenario



- Pb with entities not or badly recognized by SxPipe
- favoring longest named entities is not always the best choice !

Named Entities (cont'd)

A few classes/trees in FRMG to handle quoted NEs



Multi-word terms

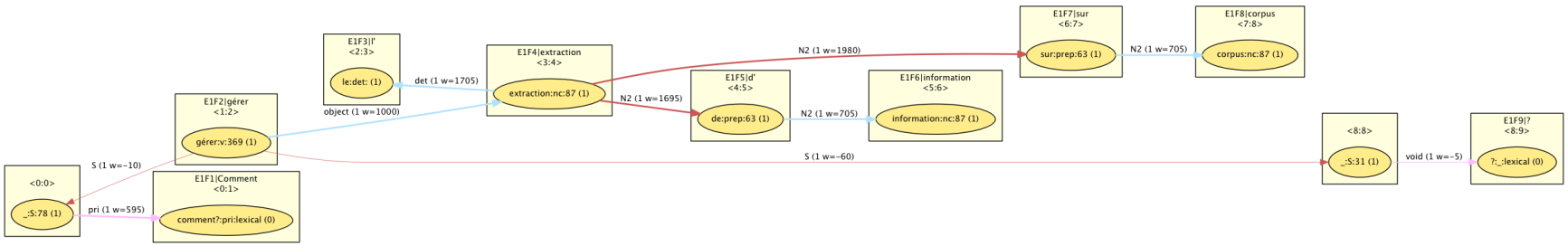
FRMG has been used to extract terms (from corpora),
They may then be injected within FRMG

- specialized domains (e.g. legal, medical)
  & big volumes (10 to 100Kterms)
- no integration in LEFFF (it is a choice !)
- could be handled by a new module in SXPIPE (as NEs)

But in practice :

- (most) terms have a regular internal syntactic structure (e.g. **N de N**)
- therefore, handled only at disamb time, with 2 rules
  - `-break_term` to penalize breaks within a term
  - `-RESTR_mod_on_term` to penalize modifiers not on term heads
- however, don't warrant we get the right internal structure for a term

Comment gérer l'extraction d'information sur corpus ?



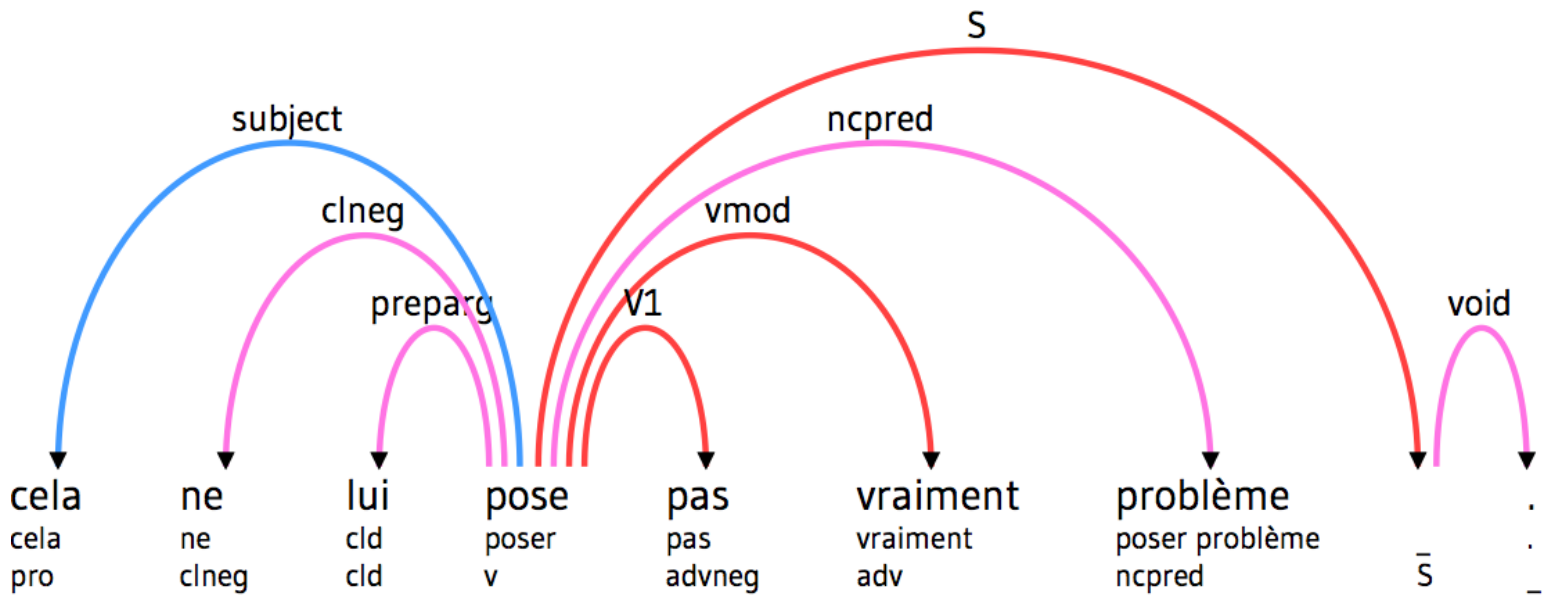
Predicative nouns and light verbs

Combination of :

- lexical information in LEFFF on predicative nouns and light verbs

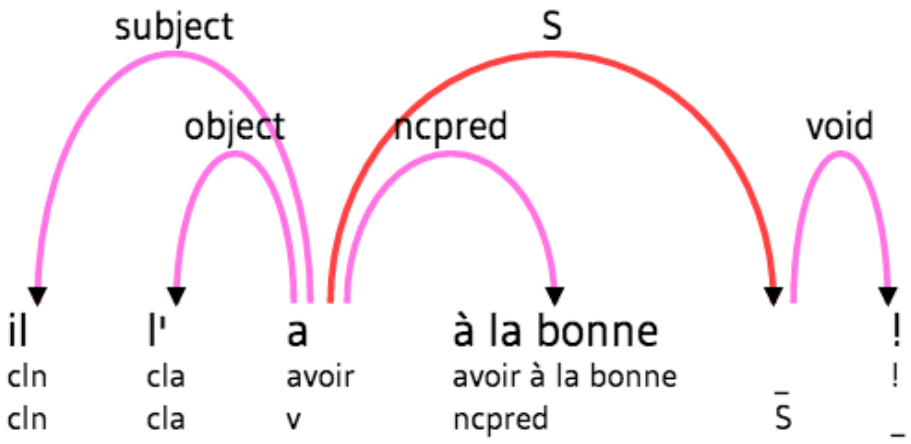
```
problème cfi [pred="problème<Suj:cln|sn,Objà:(à-sn|cld)>"
              lightverb=poser]
```

- a verbal arg `ncpred` in FRMG verbal trees (from meta-grammar)
- a transfer of subcat frame from pred. noun to light verb (at parsing time)

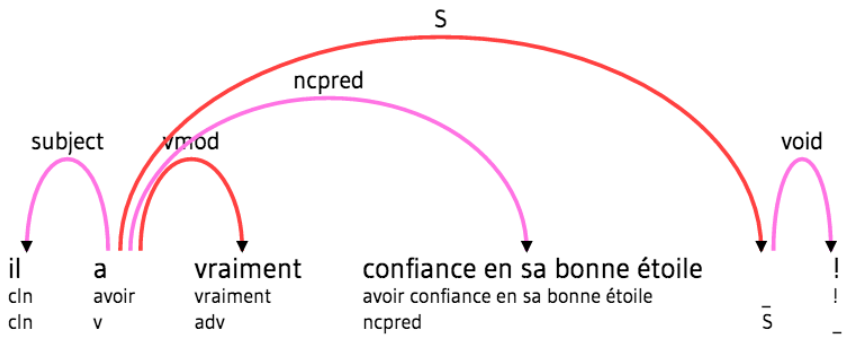


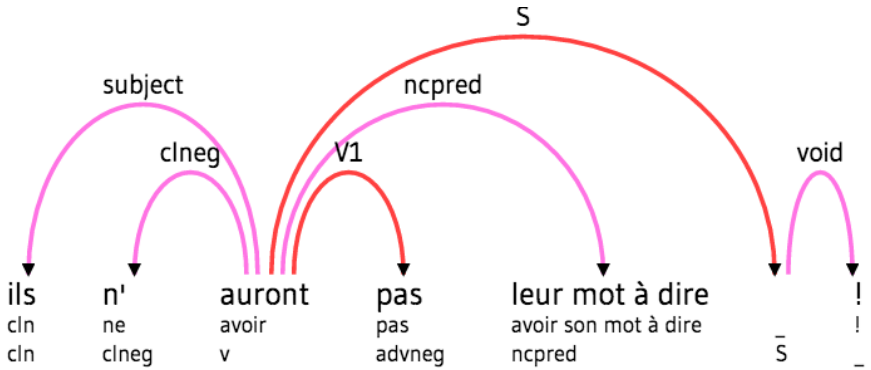
Stretching to the limits

Actually, this idea of predicative nouns extended to other categories !



Maybe going too far !



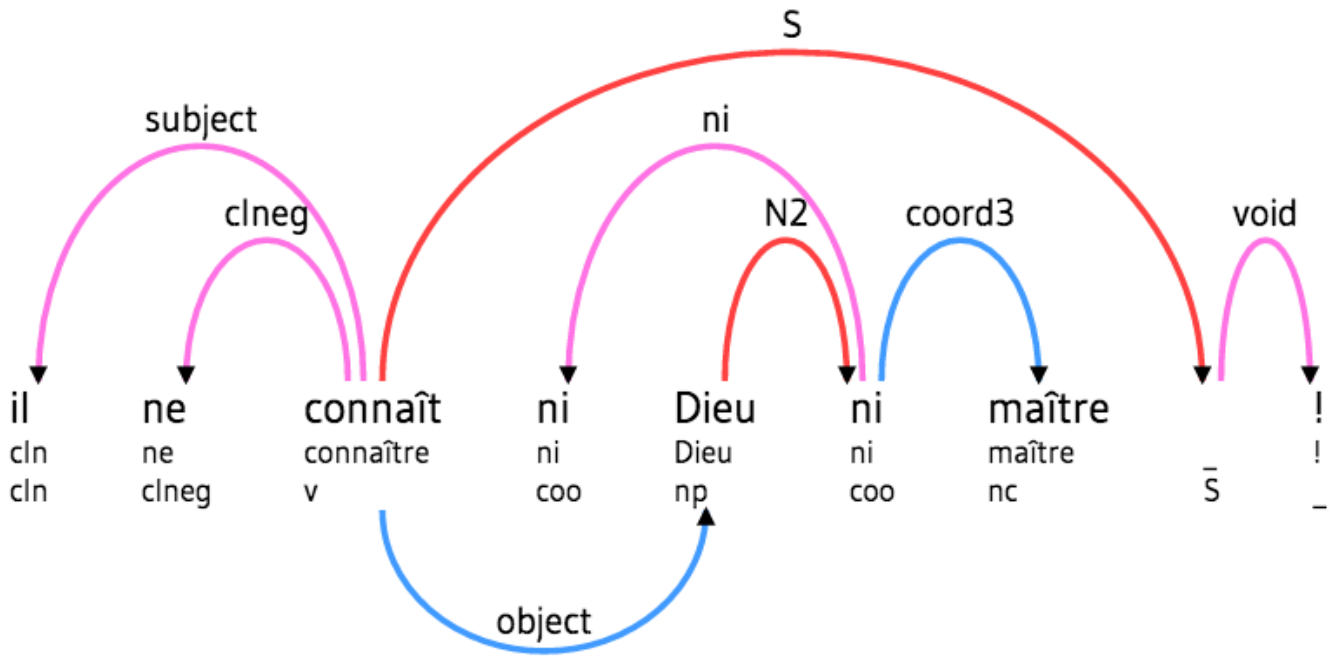


A few figures

- In LEFFF, 320 lexical entries for predicative nouns + 709 special ones
  $\Rightarrow$ many are missing !
- a few others added locally in FRMG such as « **prêter serment** »
- experiments tried on « Tables du Lexique-Grammaire » with Elsa Tolone
  30700 entries $\Rightarrow$ too many, rare cases, no probabilities
  e.g **pratiquer le yoga royal**
- Other resources welcome !

Discontinous constructions

Using anchors and co-anchors in elementary trees.



Inherit from coordination classes and apply to many categories (NP, S, AdjP, . . .)

(semi-frozen verbal) locutions

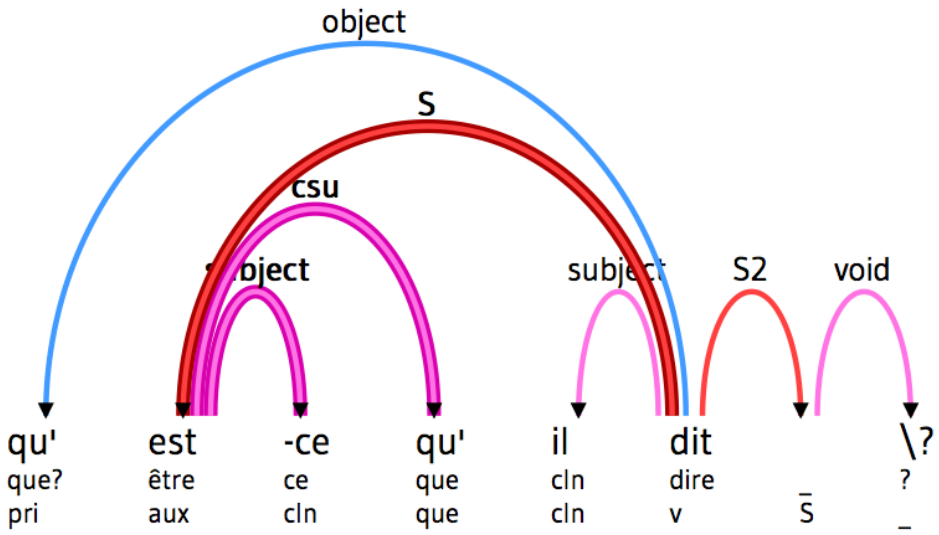
The situation is more complex for locutions that do not fit usual parts-of-speech (noun, adj, v, ...)

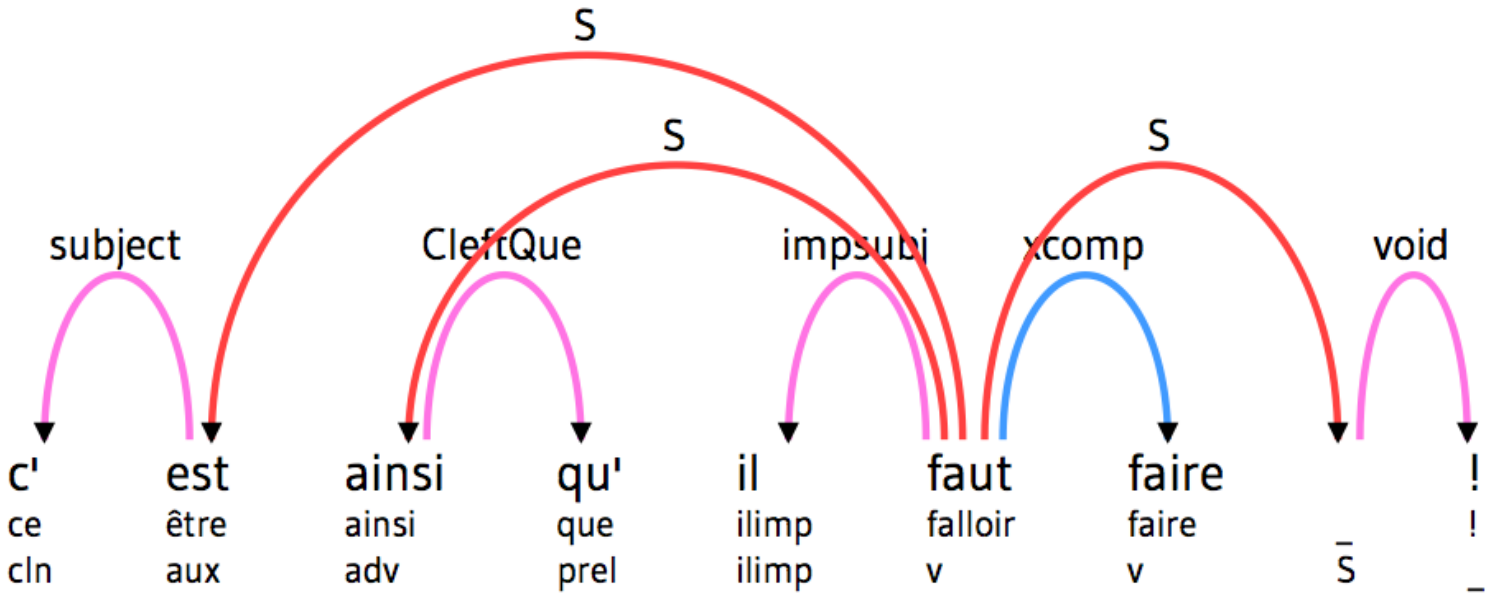
Case of **est-ce que** and **c'est X que Y** :
added a class in metagrammar ($\leadsto$ 1 tree)

- inherit from a verbal class
- + constraints (anchor=**être**, mood, imp. subject, ...)

```
class cleft_verb
{
 <: _verb_or_aux;
 node v : [cat:aux, adj: no,
           bot: [form_aux: être, diathesis: active,
                 mode: ~imperative|gerundive|infinitive
                ]];
 ...
}
```

Locution (cont'd)



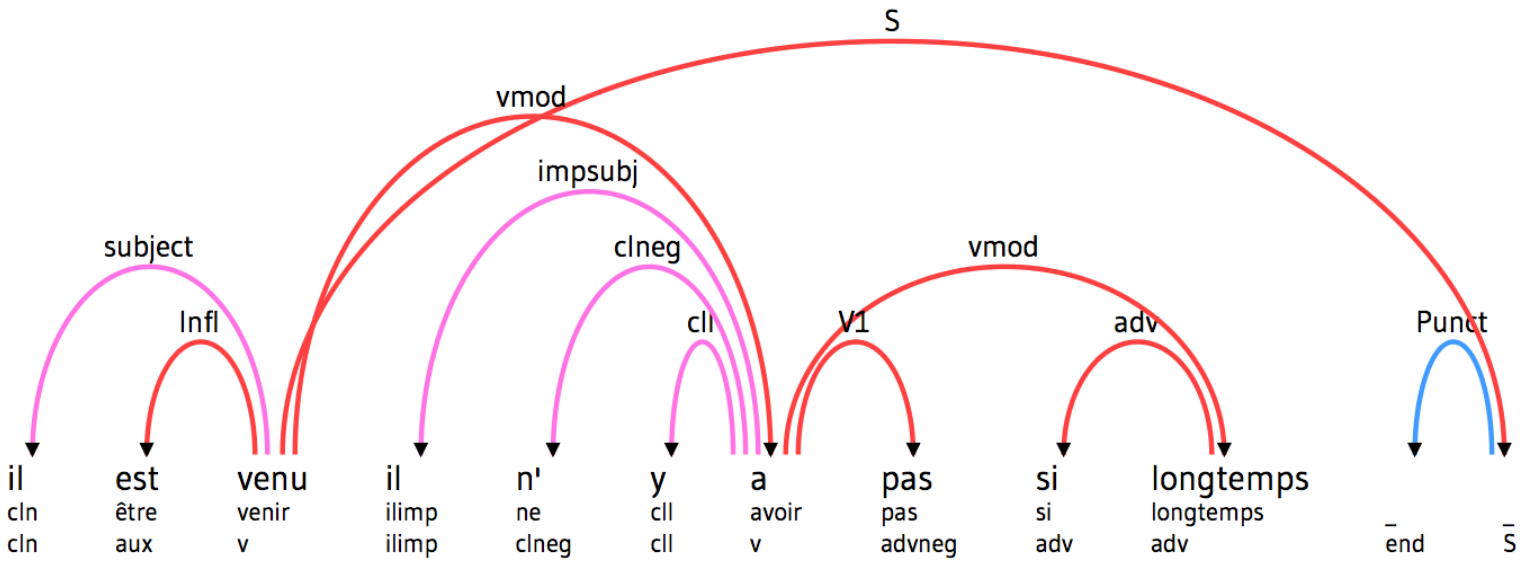


Other locutions

FRMG has a few classes for other verbal locutions

ici : **il y a** as a phrase **used** as a temporal adverbial (role)

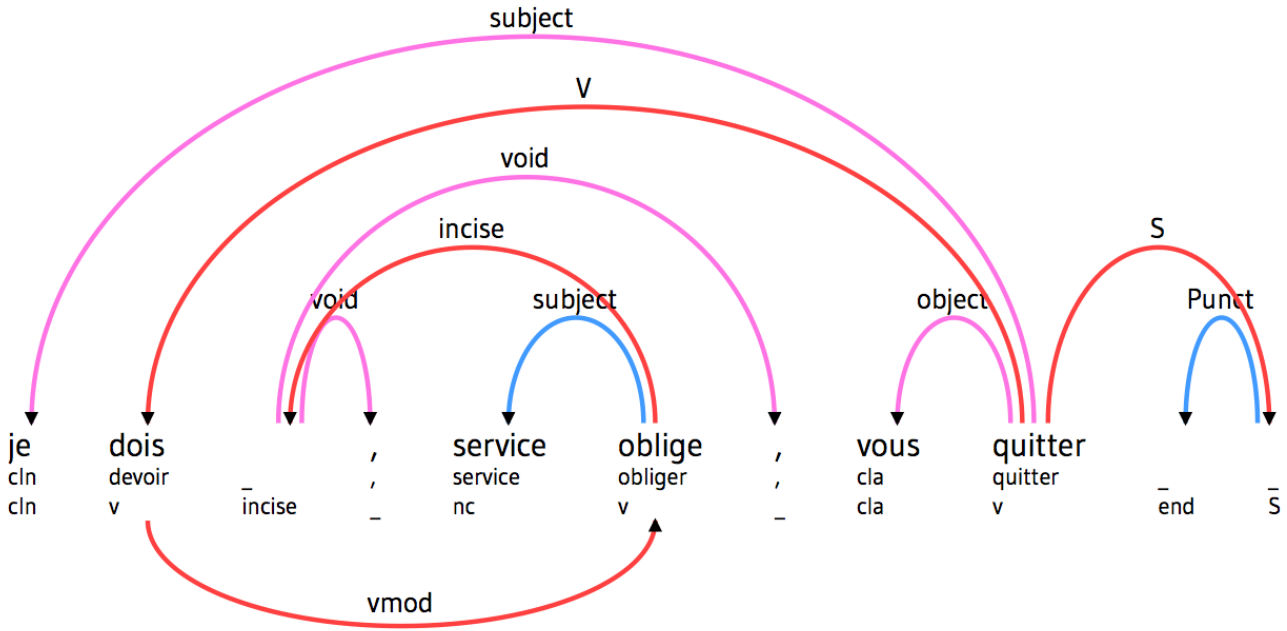
```
class verb_ilya_as_time_mod
{
    <: _verb_canonical;
    node v:[cat:v, lex: avoir];
    desc.ht.imp = value(+);
    desc.ht.extraction = value(−);
...
}
```



Other locutions (even more exotic))

Specific constraints (unsat nominal subject, . . .) and role (adverbial)

```
class verb_oblige_as_mod %% nobility obligates
{
    <: categories;
    node S: [type:std, cat: S, adj: no];
    S >> subject; S >> v; Anchor = v; subject < v;
    node subject: [ type: subst, id: subject, cat: N2, top: [sat: -]];
    node v: [ cat: v, lex: obliger, bot: [mode: indicative, person: 3]];
    node(subject).top = node(v).bot {.@ngp};
    ...
}
```



Fin de l'appartheid oblige, . . . (FTB)

Locutions : limits

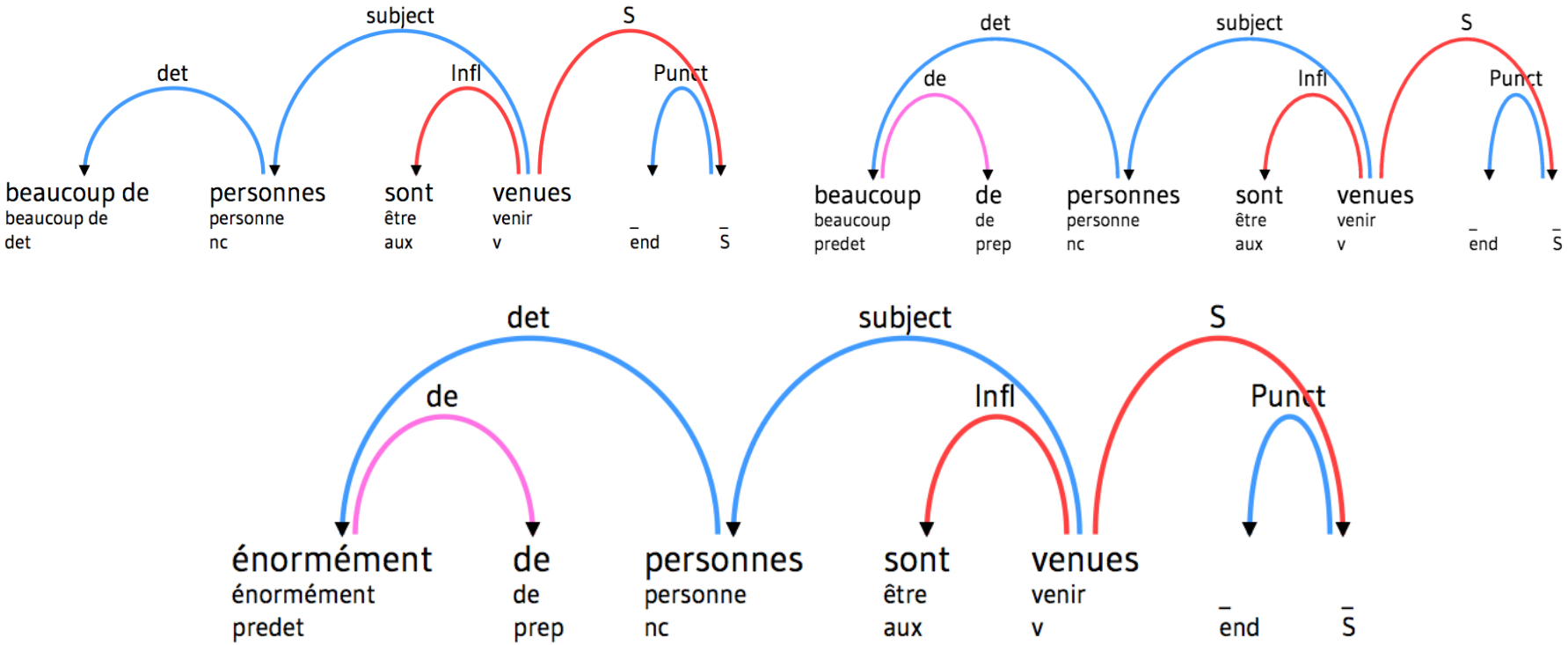
- MWEs generally not very visible in dependency trees
  may be deduced from tree names
- not elegant to add (many very specific) classes in FRMG for locutions
  best to deal with them at lexicon level
- classes/constraints not always obvious to describe
  syntactic restrictions but also semantic ones

To be done exemples :

- **n'importe quoi/où/comment/...** : il accepte **n'importe quel** travail
  actually, part of them present in LEFFF
- **je/on/nous ne sais (pas ?) qui/quoi/comment/où/...** :
  il a un **je ne sais quoi** de bizarre

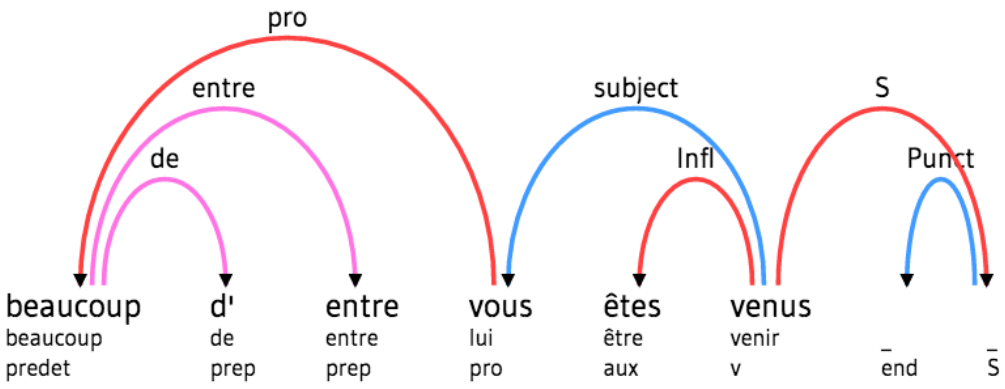
MWE : limits

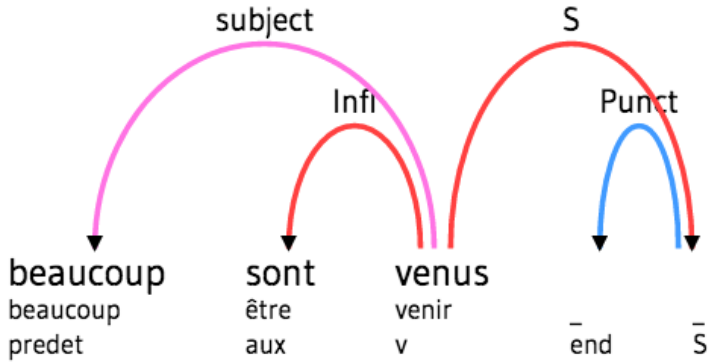
The unitary representation of MWEs in LEFFF not always the best choice :
$\Rightarrow$ mask (productive) internal syntactic structures



MWEs vs productive constructions

The notion of `predet` is productive





Multi-words preps and conjonctions resulting from a productive construction ?

afin|avant|après|au point|dans le but|de peur|... (de)/que

also, a productive construction would allow modifiers in the middle

Et **au point** parfois **de** ne plus pouvoir pénétrer dans l'appartement
Ce déploiement d'énergie est en place **afin**, souvent, **de** combler un vide intérieur.

Conversion issues

FRMG can output parses following several annotation schema :
DepXML, Passage, FTB/Conll, SPMRL, UD

However, various conventions and lists for MWEs $\Rightarrow$ conversion issues

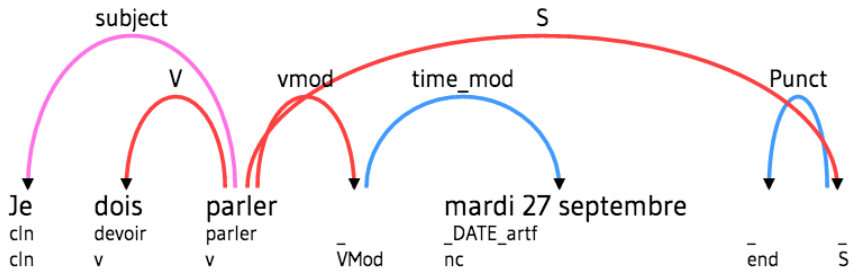
2 main cases to consider :

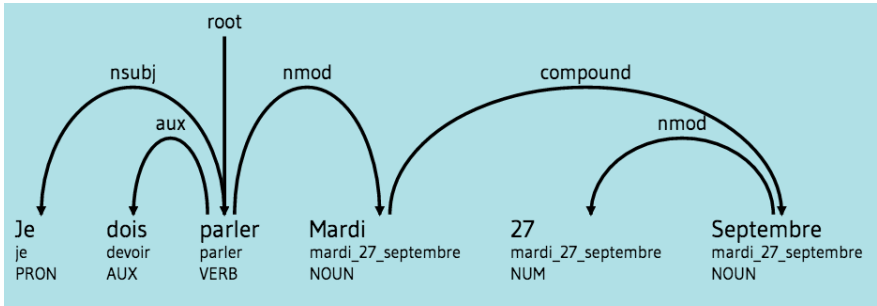
- easy one : sequence for FRMG $\rightarrow$ MWE for target schema
  $\Rightarrow$ collapse internal FRMG dependencies
- complex one : MWE for FRMG $\rightarrow$ sequence for target schema
  $\Rightarrow$ invent missing internal structure, using (schema-dependent) rules

Conversion issues (cont'd)

An exemple of conversion rule for dates targeting Universal Dependencies

```
% Mercredi 26 Novembre
udep_mwe_complex_expansion([day[],N,month[]],_,
            [head @ ('NOUN', 'nc'),
             (nmod: 1) @ ('NUM', 'adj'),
             (compound: -2) @ ('NOUN', 'nc')
            ]
) :- is_number(N).
```





Conclusion

- Several mechanisms in FRMG to handle MWEs
  tend to delay MWE processing when no syntactic impact

- Need to get something more uniform (but diversity of MWEs !)
  - better separation lexicon/(meta-)grammar (but interactions)
  - frontier between MWEs and (productive) syntactic constructions

- A possible answer : richer lexical entries of MWEs providing
  - a unitary view (when possible) as noun, adv, csu, ...
  - an internal view + flexible parts + role
    dependency structure, constraint layer, class/tree ...
  - maybe a library of specialized syntactic patterns
    (between meta-grammar & lexicon)

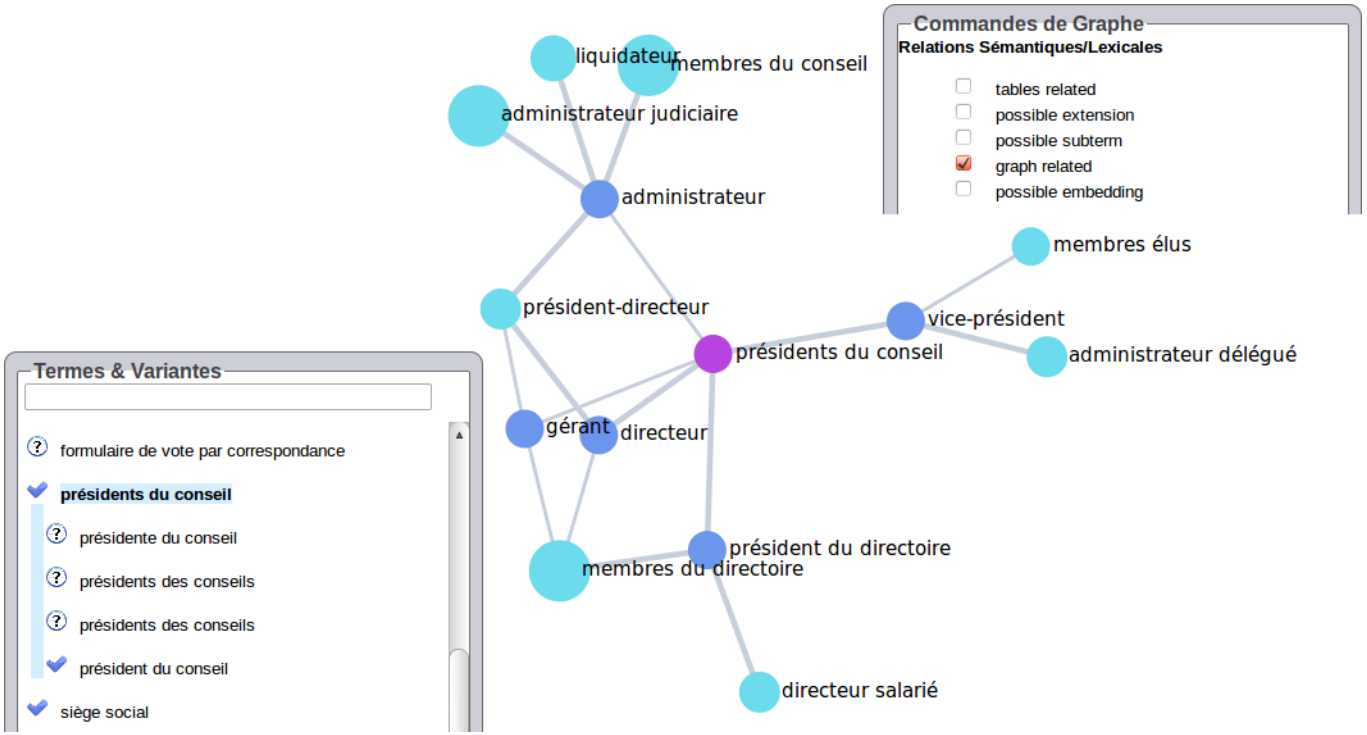
- **goal** : possibility to use FRMG's trees (rather than classes)
  instantiating/blocking configurations thanks to constraints

- however better to get representations independent from FRMG

Other themes related to MWEs

- Term extraction from large parsed corpora
  look at `http://alpage.inria.fr/Lbx` (guest/guest)

- Distributional clustering with term injection, from large parsed corpora
  again look at `http://alpage.inria.fr/Lbx` (guest/guest)



- **MAF** : Morpho-syntactic Annotation Framework, ISO standard
  clear separation between tokens and word-forms (n-m mappings)