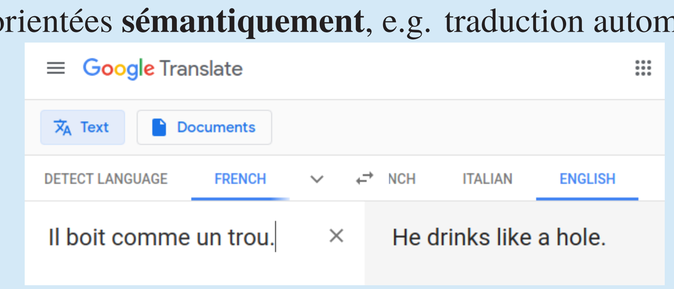
# Représentation orientée-objet d'expressions idiomatiques dans une métagrammaire

Sarah Pollet & Agata Savary & Emmanuel Schang & Anais Lefeuvre-Halftermeyer
LIFAT, Blois & LIFO, Orléans & LLL, Orléans

## Expressions polylexicales (EP: *multiword expressions*, MWEs)

*Si vous **avez** tant **besoin** de **couper l'herbe sous le pied** de quelqu'un, je vous proposerais de **vous en prendre** au **rédacteur-en-chef**, Monsieur **Jean-Marc Petit**.*

- Définition: Combinaisons de **plusieurs mots** qui possèdent des **propriétés irrégulières** au niveau du lexique, de la grammaire, de la sémantique, etc.
- Sémantique non-compositionnelle: Le sens global n'est pas déductible de manière régulière à partir des sens des composants, et des liens syntaxiques qui les relient.
  - *couper l'herbe sous le pied de quelqu'un* 'empêcher quelqu'un de réussir'
  - *s'en prendre à quelqu'un* 'prendre quelqu'un pour cible, lui attribuer une faute'
- Fréquence: Environ **20%** des mots d'un corpus français de référence appartiennent aux EP (exemple plus haut: 57%).
- Non-compositionalité:
  - Méthodes informatiques sont **compositionnelles**
    - Phénomènes complexes sont décomposés en problèmes plus simples.
    - Sous problèmes reçoivent des solutions autonomes, qui sont ensuites composées pour fournir solutions globales.
  - EP sont **sémantiquement non-compositionnelles**, donc posent problème pour le tâches du **TAL** orientées **sémantiquement**, e.g. traduction automatique.




## Figement

Une EP est **moins flexible** (variable) qu'une construction régulière de la même structure syntaxique.

| Structure | Construction régulière | Expression polylexicale |
|---|---|---|
| N Adj | *livre bleu* | *cordon bleu* 'excellent cuisinier' |
| V Det N | *manger des salades*<br>*faire la soupe* | *raconter des salades* 'mentir'<br>*faire la tête* 'bouder' |
| V Prep Det N | *être sur son siège*<br>*cuisiner au vinaigre* | *être sur son trente-et-un* 'être très élégant'<br>*tourner au vinaigre* 's'orienter vers la dispute' |
| Adj Coord Adj | *rouge et bleu* | *noir et blanc* 'nuances de gris' |
| Det N Aux V | *les patates sont cuites* | *les carottes sont cuites* 'la situation est compromise' |

| Construction régulière | Expression polylexicale | Propriété |
|---|---|---|
| *livre bleu* $\approx^{1}$ *livre cyan* $\approx$ *bouquin bleu* | *cordon bleu* vs. *#cordon cyan* vs. *#corde bleue* | Figement lexical |
| *manger des salades* $\approx$ *manger une salade* | *raconter des salades* vs. *#raconter une salade* | Figement morphologique |
| *elle est sur son siège* $\approx$ *elle est sur mon siège* | *elle est sur son trente-et-un* vs. *#elle est sur mon trente-et-un* | Figement morphosyntaxique |
| *cuisiner au vinaigre* $\approx$ *cuisiner au vinaigre de vin* | *tourner au vinaigre* vs. *#tourner au vinaigre de vin* | Figement syntaxique |
| *il a fait la soupe* $\approx$ *la soupe a été faite par lui* | *il a fait la tête* vs. *#la tête a été faite par lui* | |
| *les patates sont cuites* $\approx$ *nous avons cuit les patates* | *les carottes sont cuites* vs. *#on a cuit les carottes* | |
| *un texte en rouge et bleu* $\approx$ *un texte en bleu et rouge* | *une photo en noir et blanc* vs. *#une photo en blanc et noir* | |

'$\approx$': glissement de sens prévisible à partir du remplacement lexical

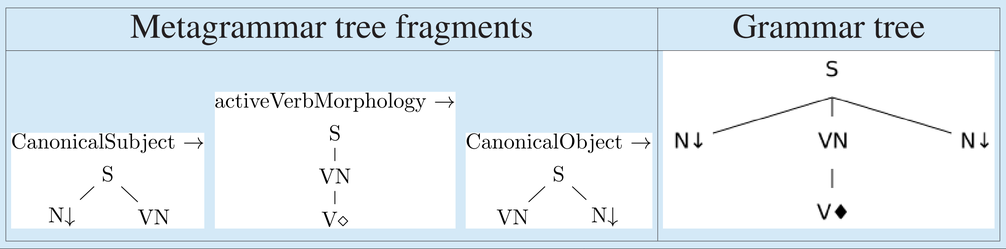
Le figement des EP est une question d'**échelle** et non pas une propriété binaire [Gross(1986), Gross(1988)]

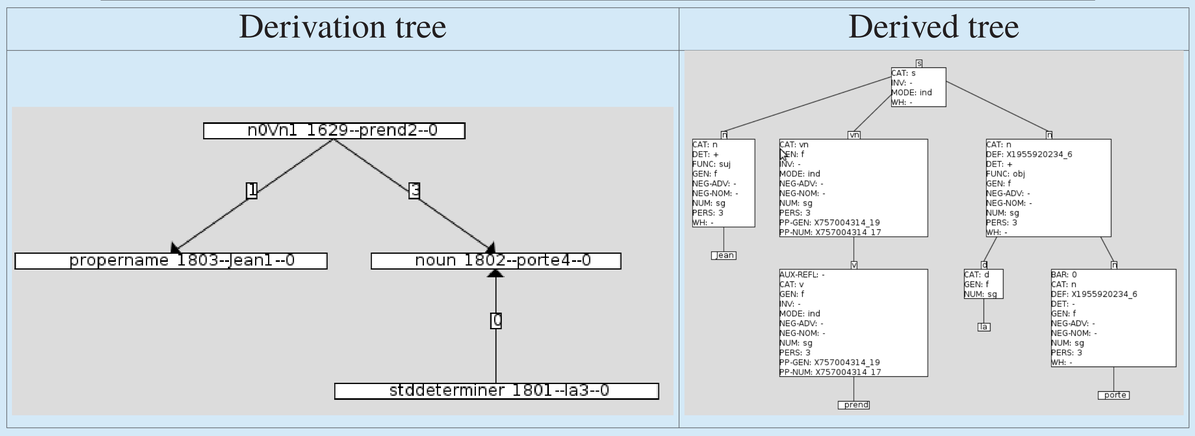
| $N_0V(DetN)_1$ expr. | Free subject | Free verb | Free object | Verb red. | Verb infl. | Noun infl. | Noun modif. | Passive | Det. altern. |
|---|---|---|---|---|---|---|---|---|---|
| *N prend la pomme* | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *N prend une décision* | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *N tourne la page* | ✓ | | | | ✓ | ? | ✓ | ✓ | ? |
| *N prend la porte* | ✓ | | | | ✓ | | | | |

## Metagrammatical description

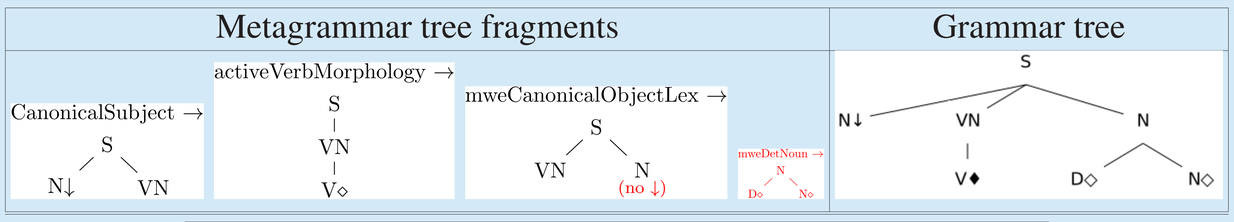
- **XMG** [Crabbé et al.(2013), Petitjean et al.(2016)]
  - a language
    - **object-oriented** – objects, classes, inheritance
    - declarative – grammaticality = **constraints** rather than procedures
    - notationally **expressive** - modularity, inheritance, conjunction/disjunction of tree fragments, namespaces
    - **extensible** to new dimensions (semantics, frames etc.), formalisms, . . .
  - a metagrammar **compiler** (for each target language, here FS-LTAG) – constraint solver: produces minimal tree models respecting the constraints
- **FrenchTAG** – French XMG metagrammar [Crabbé et al.(2013)], XMG implementation of the syntactic Tree Adjoining Grammar (TAG) of French [Abeillé(2002)]
  - 285 XMG classes, 96 families, compiled into 9045 TAG trees
  - toy lexicon of 555 lexemes, including 248 verbs
  - covers **literal** readings (by compositionality) but **not idiomatic** readings

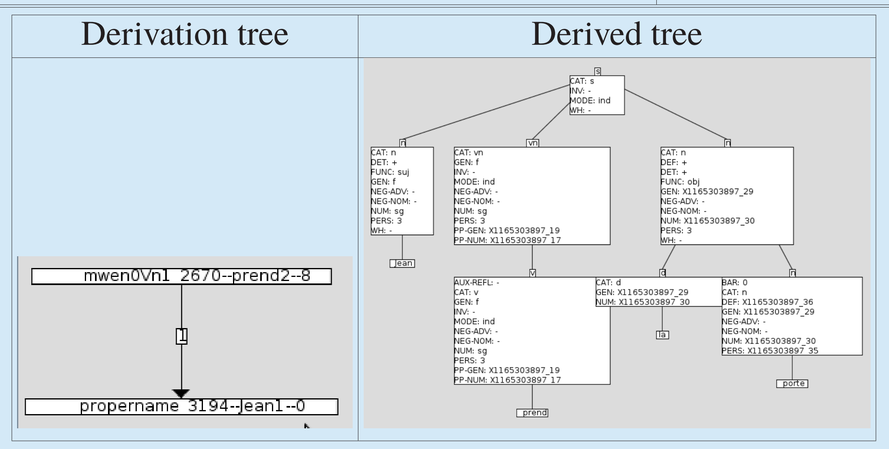
## Parsing a literal reading: (*Jean prend la porte*)




## Parsing an indiomatic reading: (*Jean prend la porte*)




## Lexicon: morphology and lemmas

```
class Jean
{
  <morpho> {
    morph <- "Jean";
    lemma <- "jean";
    cat   <- n
  }
}
```

```
class prend
{
  <morpho> {
    morph <- "prend";
    lemma <- "prendre";
    cat   <- v
  }
}
```

```
class la
{
  <morpho> {
    morph <- "la";
    lemma <- "le";
    cat   <- d;
    gen <- f
  }
}
```

```
class porte
{
  <morpho> {
    morph <- "porte";
    lemma <- "porte";
    cat   <- n
  }
}
```

propename → N◇

```
class LemmeJean
{
  <lemma> {
    entry <- "jean";
    cat   <- n;
    fam   <- propername
  }
}
```

```
class LemmePrendre
{
  <lemma> {
    entry <- "prendre";
    cat   <- v;
    fam   <- n0Vn1
  }
}
```

```
class LemmeLe
{
  <lemma> {
    entry <- "le";
    cat   <- d;
    fam   <- stddeterminer
  }
}
```

```
class LemmePorte
{
  <lemma> {
    entry <- "porte";
    cat   <- n;
    fam   <- noun
  }
}
```

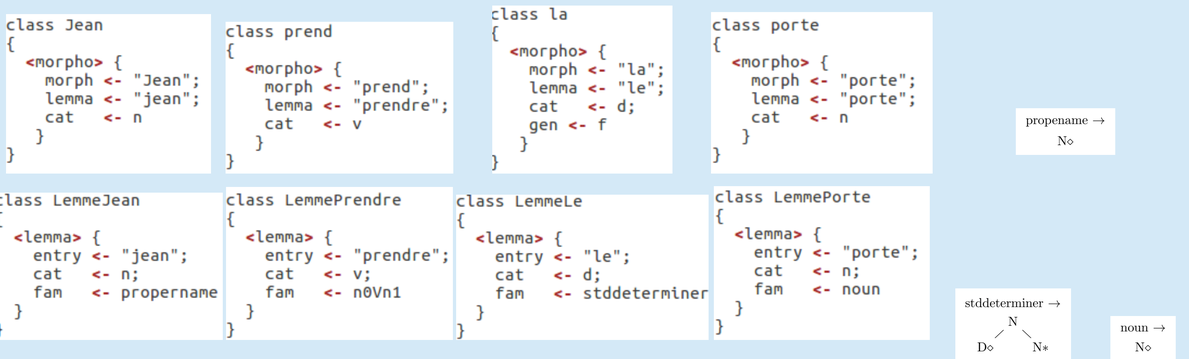
stddeterminer → N (D◇, N∗)

noun → N◇

```
class mweLemmePrendreLaPorte
{
  <lemma> {
    entry <- "prendre";
    cat   <- v;
    fam   <-    mwen0Vn1;
    filter dia = active;
    filter subj = free;
    filter obj = lexicalized;
    filter objstruct = lexDetLexN;
    coanchor ObjDetNode -> "la"/d;
    coanchor ObjNode -> "porte"/n;
    equation ObjNode -> gen=f;
    equation ObjNode -> num=sg;
    equation ObjNode -> modifiable=-;
    filter objtype = canonical
  }
}
```

## ICVL internship: Developing a real size lexicon for FrenchTAG

- Original FrenchTAG lexicon: toy lexicon of 555 lexemes, including 248 verbs
- French lexical resources:
  - **Lefff** - Lexique des Formes Fléchies du Français [Sagot, 2010]: 110 000 lemmas, 520 000 formes




  - **Lexicom-Grammar** [Gross 1975, Tolone 2011] - 40,000 EP

| <ID> | N0 =: Nhum | N0 =: N-hum | Ppv =: se figé | <ENT>Ppv | <ENT>V | N0 V | <ENT>Det1 | <ENT>C1 | C1 =: Npc | [passif] | Exemple |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 142 | + | - | - | <E> | boire | - | un | verre | - | + | §boire§un verre |
| 1153 | + | - | - | <E> | prendre | - | un | bain de foule | - | - | §prendre§un bain de foule |

- Objectives: convert Lefff and the Lexicon-Grammar to XMG
- Outcome:
  - 
  - XMG lexicon with **112,000 lemmas** and their inflected forms