

# Towards a MWE-aware French metagrammar in XMG

Agata Savary <sup>1,2</sup>, Simon Petitjean <sup>2</sup>, Laura Kallmeyer <sup>2</sup>,  
Timm Lichte <sup>2</sup>

<sup>1</sup>Université de Tours, France

<sup>2</sup>University of Düsseldorf, Germany

TreeGraSP meeting  
21 February 2018  
Düsseldorf

## Multiword expressions

- Word combinations, which exhibit lexical, syntactic, semantic, pragmatic and/or statistical **irregularities**.
- Examples: *all of a sudden*, *a hot dog*, *to pay a visit*, *to pull one's leg*
- Encompass **heterogeneous objects**: idioms, compounds, light verb constructions, rhetorical figures, institutionalised phrases or named entities
- Pervasive feature: **non-compositional semantics** - the meaning of an MWE cannot be deduced from the meanings of its components, and from its syntactic structure, in a way deemed regular for the given language.

# Lexical, morphological and syntactic properties of MWEs

## Types of properties

- **Defective** property – excludes a literal interpretation of the MWE, e.g.:
  - Defective agreement: *a cross-roads*
- **Restrictive** property – reduces the number of possible surface realizations of the MWE with respect to the literal reading, e.g.:
  - Restrictive lexical selection: *to make ends meet* vs. *#to make edges come together*
  - Restrictive agreement: *I cross my fingers* vs. *#I cross his fingers*
  - Restrictive diathesis: *the die is cast* vs. *#one casts the die*
  - Restrictive modification: *he lives a life of luxury* vs. *#he lives a life*
  - Restrictive dependencies between determiners and modifiers: *j'ai envie de le faire* 'I have desire to do it' ⇒ 'I feel like doing it', *j'ai une envie folle de le faire* 'I have a desire crazy to do it' ⇒ 'I terribly feel like doing it'

# Scale-wise regularity

More regular ( $\succ$ ) = admitted by more objects (in a set)

- sample set: English **Subj-Verb-Obj** expressions (*John pulled the door*)
- “allow any head verb”  $\succ$  “allow only the verb *kick*”
- “allow passive”  $\succ$  “prohibit passive”
- “allow a possessive determiner”
  - John pushed the/my door*
  - ∧ “impose a possessive determiner”
    - John broke his/our fall* ‘John made his/our fall less forceful’
    - ∧ “impose a possessive agreeing with Subj”
      - John crossed his fingers* ‘John hoped for good luck’
      - John held his tongue* ‘John refrained from expressing his view’

Idiosyncratic = irregular (shared by no two objects)

- Usually only the restrictive lexical selection is truly idiosyncratic (except in polysemous MWEs: *go on* ‘continue/happen’)

# Lexical encoding of MWEs

## Linguistic tradition of MWE encoding

- Lexicon-grammar [Gross(1986)]
- Explanatory Combinatorial Dictionary [Mel'čuk *et al.*(1988)]
- Some NLP applications:
  - LG: [Hathout and Namer(1997b), Hathout and Namer(1997a), Hathout and Namer(1998), Gardent *et al.*(2005), Gardent *et al.*(2006), Constant and Tolone(2010), Laporte *et al.*(2013), Tolone and Sagot(2011)]
  - DEC: [Apresian *et al.*(2003), Lambrey and Lareau(2015)]

# Lexical encoding of MWEs

## NLP-oriented encoding

- Dozen formalisms for continuous MWEs (7 languages) [Savary(2008)]
- Verbal MWEs:
  - morphosyntactic databases (NL) [Grégoire(2010)], (HE) [Al-Haj *et al.*(2014)]
  - valence dictionaries (CS) [Hajič *et al.*(2003)] (PL) [Przepiórkowski *et al.*(2014)]
  - ontological approaches with semantic calculus: (EN) [Marjorie McShane and Beale(2005)]

## Redundancy and implicitness issues

- Capturing regularity: inflection codes [Savary(2009)], equivalence classes [Grégoire(2010)], macros [Przepiórkowski *et al.*(2014)]
- Implicit interface with a "regular" grammar despite its crucial role in the formalism [Grégoire(2010)], [Przepiórkowski *et al.*(2014)], [Marjorie McShane and Beale(2005)]

## MWEs in LTAGs

[Abeillé and Schabes(1989)]

- Representing **discontinuities** (cf. *extended domain of locality*)
  - discontinuities in the internal structure of a MWE  $\Rightarrow$  visible in ETs, handled by **substitution**

*to take something with a pinch of salt*
  - insertion of adjuncts  $\Rightarrow$  handled by **adjunction**

*a whole bunch of NP*
- Dependencies between arguments at **different depths** in the ETs are naturally expressed
 

*She pulled her/#its socks up*

# Redundancy in grammar encoding

## Redundancy in MWEs (and “regular” structures)

- properties shared by “regular” structures and by MWEs (e.g. passivisation, extraction etc.),
- properties shared by many MWEs (e.g. poss.-subj. agreement),
- properties of different degrees of regularity co-occur in each MWE.

## Redundancy in a TAG grammar

- elementary trees of a lexicalized grammar are very numerous (hundreds or thousands of trees),
- elementary trees share tree fragments and their properties.



# Motivation

## Objectives

- avoid redundancy in MWE encoding
- abstract away (as much as possible) from the actual grammatical formalism

## Object-oriented encoding

- represent the shared tree fragments and properties as **classes**
- combine these classes into complete minimal structures
- class hierarchy:
  - more general properties – encoded in upper upper classes
  - less general ones – encoded in lower classes (which inherit from the upper ones)

# XMG [Crabbé et al.(2013), Petitjean et al.(2016)]

- a language
  - declarative – grammaticality is defined in terms of **constraints** rather than procedures
  - notationally **expressive** - modularity, inheritance, conjunction/disjunction of tree fragments, namespaces
  - **extensible** to new dimensions (semantics, frames etc.), formalisms (IG, etc.), linguistic principles (e.g. clitic ordering)
- a metagrammar **compiler** (for each tager 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 TAG grammar of French by [Abeillé(2002)]
  - 285 XMG classes, 96 families (classes assigned to lexemes), compiled into 9045 TAG trees
  - toy lexicon of 555 lexemes, including 248 verbs
- SemTag – extension of FrenchTAG with a (compositional) semantic dimension

### Example

- *Jean prend la porte* 'John takes the door' ⇒ 'John leaves because he is forced to'
- XMG covers **literal** readings (by compositionality)
- XMG does **not** cover **idiomatic** readings

# Morphology (simplified)

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

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

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

```
class il
{
  <morpho> {
    morph <- "il";
    lemma <- "il";
    cat   <- cl
  }
}
```

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

```
class laClitic
{
  <morpho> {
    morph <- "la";
    lemma <- "le";
    cat   <- cl
  }
}
```

# Lemmas

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

```
class LemmeIl
{
  <lemma> {
    entry <- "il";
    cat   <- cl;
    fam   <- CliticT
  }
}
```

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

```
class LemmeLeClitic
{
  <lemma> {
    entry <- "le";
    cat   <- cl;
    fam   <- CliticT
  }
}
```

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

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

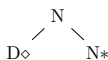
## Trivial classes

propname →  
N◊

noun →  
N◊

CliticT →  
CL◊

stddeterminer →



# From metagrammar to parsing: **n0Vn1** (*Jean prend la porte*)

## Metagrammar tree fragments inherited by n0Vn1

CanonicalSubject →



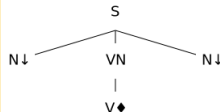
activeVerbMorphology →



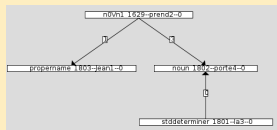
CanonicalObject →



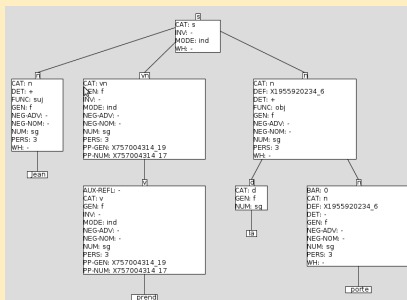
## Grammar tree



## Derivation tree



## Derived tree



# From metagrammar to parsing: **n0Vn1** (*Il prend la porte*)

## Metagrammar tree fragments inherited by n0Vn1

CliticSubject →



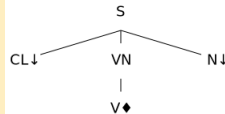
activeVerbMorphology →



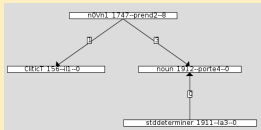
CanonicalObject →



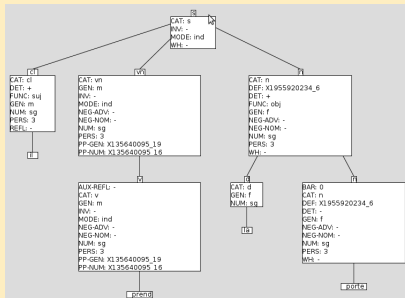
## Grammar tree



## Derivation tree



## Derived tree



# From metagrammar to parsing: n0Vn1 (*Jean la prend*)

## Metagrammar tree fragments inherited by n0Vn1

## Grammar tree

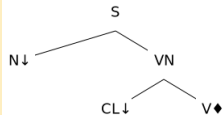
CanonicalSubject →



activeVerbMorphology →



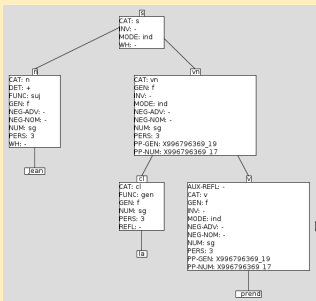
CliticObject →



## Derivation tree



## Derived tree





# XMG classes

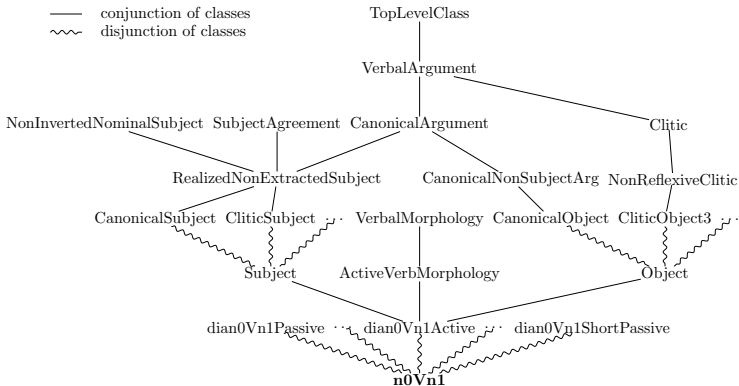
```
class Subject
{
  CanonicalSubject[]
  | CliticSubject[]
  | RelativeSubject[]
  % Other alternatives omitted
}
```

```
class Object
{
  CanonicalObject[]
  | CliticObjectII[]
  | CliticObject3[]
  | RelativeObject[]
  % Other alternatives omitted
}
```

```
class dian0Vn1Active{
  Subject[];
  activeVerbMorphology[];
  Object[]
}
```

```
class n0Vn1{
  dian0Vn1Active[]
  | dian0Vn1Passive[]
  | dian0Vn1ShortPassive[]
  % Other alternatives omitted
}
```

# Class hierarchy



## MWEs have more or less regular properties

*prendre la porte* 'to take the door' ⇒ 'to leave when one is forced to'

- (More) **regular** features:

- The subject is free and agrees with the verb:

*Jean/il/elle prend la porte* 'Jean takes the door' ⇒ 'Jean lives ...'

*Jean, que nous ne voulons pas ici, prend la porte* 'Jean, whom we do not want here, takes the door' ⇒ 'Jean, whom ..., leaves ...'

- The verb inflects freely: *Prend la porte!* 'Take the door!' ⇒ 'Leave!'

- (More) **idiosyncratic** features:

- The object is lexicalized: *#Jean prend la sortie* 'Jean takes the exit'

- The object cannot be:

cliticised: *#Jean la prend* 'Jean takes it'

extracted: *#La porte que Jean prend* 'The door that Jean takes'

modified: *#Jean prend la grande porte* 'Jean takes the big door'

- The verb cannot be passivized: *#La porte est prise par Jean* 'The door is taken by Jean'

# Adding MWEs to the metagrammar

## Strategy

- add **lexical entries** for MWEs with **co-anchors** and interface **filters**,
- **reuse** existing tree fragments for the (more) regular properties, decorate them with **interface features**,
- **create** new tree fragments for **lexicalized arguments** of various syntactic structures

## MWE lemmas with **co-anchors** and **filters**

```

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
  }
}

```

## Previous tree fragments decorated with **interface features**

```
class mweSubject
{
  Subject[] *= [subj=free]
  | mweSubjectLexStruct[] *= [subj=lexicalized]
}

class mweObject
{
  Object[] *= [obj=free]
  | mweObjectLexStruct[] *= [obj=lexicalized]
}

class mwedian@Vn1Active
{
  mweSubject[];
  activeVerbMorphology[];
  mweObject[]
}

class mwen@Vn1{
  mwedian@Vn1Active[] *=[dia = active]
  | mwedian@Vn1Passive[] *=[dia = passive, passivetype = full]
  | mwedian@Vn1ShortPassive[] *=[dia = passive, passivetype = short]
  % Other alternatives omitted
}
```

## New XMG classes for lexicalized arguments

```

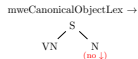
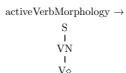
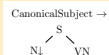
class mweLexDetLexNoun[idDetNode,idNounNode]
import mweNoDetLexNoun[idNounNode]
declare ?xLexDet ?fY ?fZ
{
  <syn>{
    node xLexNPtop[bot = [num = ?fY, gen = ?fZ]]{
      node xLexDet (color=red,mark=coanchor,name=idDetNode)[cat = d,bot=[num = ?
      node xLexNHead[cat = n,top=[num = ?fY,gen = ?fZ]]
    }
  }
}

class mweObjectLexStruct
import mweObjectLex[]
declare ?lexNP
{
  {
    ?lexNP = mweLexDetLexNoun[ObjDetNode,ObjNode] *= [objstruct=lexDetLexN
    | ?lexNP = mweNoDetLexNoun[ObjNode] *= [objstruct=lexN]
    | ?lexNP = mweLexNounLexAdj[ObjNode,ObjAdjNode] *= [objstruct=lexNLexAdj]
  };
  ?LexObj = ?lexNP.xLexNPtop
}

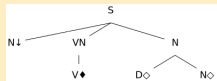
```

# From metagrammar to parsing: **mwen0Vn1** (*Jean prend la porte*)

## Tree fragments inherited by mwen0Vn1



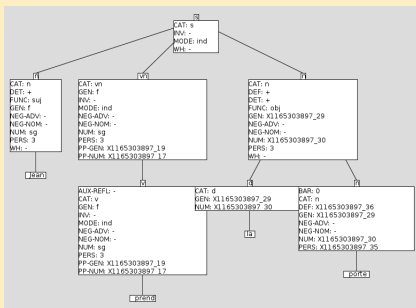
## Grammar tree



## Derivation tree



## Derived tree





# From metagrammar to parsing: mwen0Vn1

(Il prend la porte)

### Tree fragments inherited by mwen0Vn1

CliticSubject →

```

      S
     / \
    CL↓  VN
          
```

activeVerbMorphology →

```

      S
     / |
    VN |
     | |
    V◊
          
```

mweCanonicalObjectLex →

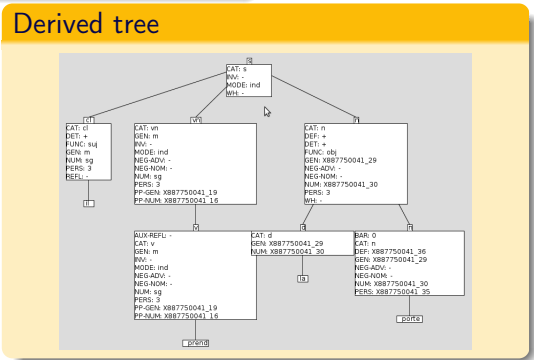
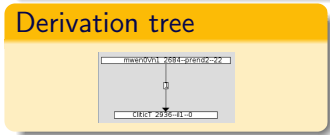
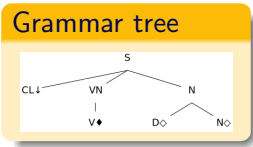
```

      S
     / \
    VN  N
         (no ↓)
          
```

mweDetNoun →

```

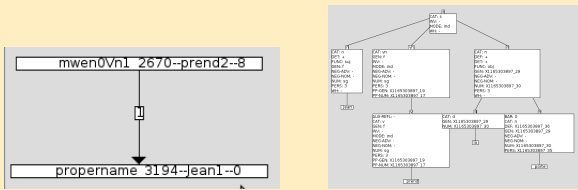
      N
     / \
    D◊  N◊
          
```



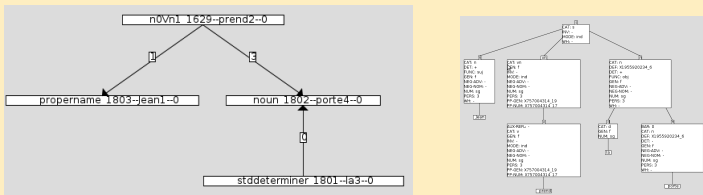


# Two readings: *Jean prend la porte*

## Idiomatic reading

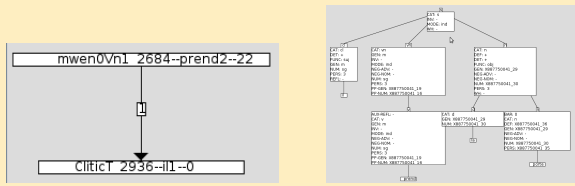


## Compositional reading

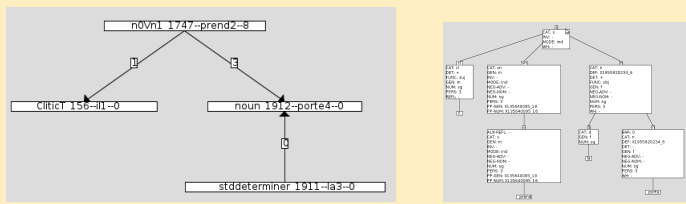


# Two readings: *Il prend la porte*

## Idiomatic reading



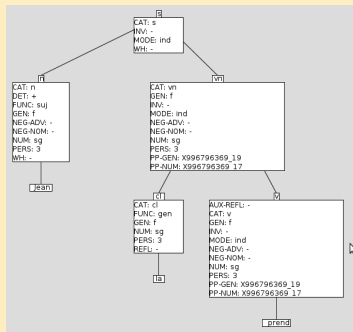
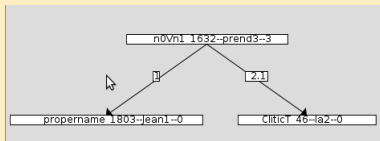
## Compositional reading



One reading: *Jean la prend*

No idiomatic reading

Compositional reading



# Conclusions and future work

## Advantages from the XMG encoding of MWEs

- **Explicit** declarative encoding of the properties of MWEs, both (more) regular and (more) idiosyncratic
- **Scale-wise modeling**: regularity/idiosyncrasy are not modeled as binary phenomena
- **Non-redundancy**: properties shared by objects (MWEs or compositional structures) are uniquely described and shared
- Direct integration into a **grammar**

## Future work

- **Encode more** MWEs and properties
- Enhance the XMG formalism and the parser (allow: lemmas as co-anchors, lexicalized non-adjunction nodes)
- Add a **semantic** dimension based on **frames**

# Bibliography I



Abeillé, A. and Schabes, Y. (1989).

Parsing idioms in lexicalized tags.

In H. L. Somers and M. M. Wood, eds., *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, pp. 1–9. The Association for Computer Linguistics.



Abeillé, A. (2002).

*Une grammaire électronique du français*. CNRS Editions.



Al-Haj, H., Itai, A., and Wintner, S. (2014).

Lexical Representation of Multiword Expressions in Morphologically-complex Languages.

*International Journal of Lexicography*, 27(2), 130–170.



Apresian, J., Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V., and Tsinman, L. (2003).

ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT.

In *First International Conference on Meaning-Text Theory (MTT 2003), Paris, Ecole Normale Supérieure*.



Constant, M. and Tolone, E. (2010).

A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables.

In M. D. Gioia, ed., *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie*, pp. 79–93. Aracne.

ISBN 978-88-548-3166-7.



Crabbé, B., Duchier, D., Gardent, C., Roux, J. L., and Parmentier, Y. (2013).

XMG: extensible metagrammar.

*Computational Linguistics*, 39(3), 591–629.

# Bibliography II



Gardent, C., Guillaume, B., Perrier, G., and Falk, I. (2005).

Maurice Gross' grammar lexicon and Natural Language Processing.

In Z. Vetulani, ed., *2nd Language & Technology Conference, April 21-23, 2005, Poznań, Poland (LTC'05)*, pp. 120–123.



Gardent, C., Guillaume, B., Perrier, G., and Falk, I. (2006).

Extraction d'information de sous-catégorisation à partir des tables du LADL.

In *Traitement Automatique de la Langue Naturelle - TALN 2006*, Leuven/Belgique.



Grégoire, N. (2010).

DuELME: a Dutch electronic lexicon of multiword expressions.

*Language Resources and Evaluation*, 44(1-2).



Gross, M. (1986).

Lexicon-grammar: The Representation of Compound Words.

In *Proceedings of the 11th Conference on Computational Linguistics*, pp. 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.



Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003).

PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation.

In J. Nivre and E. Hinrichs, eds., *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Norway.



# Bibliography III



Hathout, N. and Namer, F. (1997a).

Génération (semi)-automatique de ressources lexicales réutilisables à grande échelle. Conversion des tables du LADL.

In *Actes des 1res JST FRANCIL*, Avignon. AUPELF-UREF.



Hathout, N. and Namer, F. (1997b).

(Semi-)automatic generation of ALEP analysis lexicon.

In *Proceedings of the 3rd ALEP User Group Workshop*, Saarbrücken.



Hathout, N. and Namer, F. (1998).

Automatic construction and validation of French large lexical resources: Reuse of verb theoretical linguistic descriptions.

In *Proceedings of the First International Conference on Language Resources and Evaluation*, pp. 627–636, Granada. ELRA.



Lambrey, F. and Lareau, F. (2015).

Le traitement des collocations en génération de texte multilingue.

In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, pp. 579–585, Caen, France. Association pour le Traitement Automatique des Langues.



Laporte, E., Tolone, E., and Constant, M. (2013).

Conversion of Lexicon-Grammar tables to LMF. Application to French.

In G. Francopoulo, ed., *LMF. Lexical Markup Framework*, pp. 157–187. ISTE - Wiley.

# Bibliography IV



Lichte, T., Parmentier, Y., Petitjean, S., Savary, A., and Waszczuk, J. (2016).

Separating the regular from the idiosyncratic: A constraint-based lexical encoding of MWEs using XMG.

<http://typo.uni-konstanz.de/parseme/index.php/2-general/156-selected-posters-struga-7-8-april-2016>.



Marjorie McShane, S. N. and Beale, S. (2005).

The description and processing of multiword expressions in ontosem.

Working Paper 07-05, Institute for Language and Information Technologies University of Maryland Baltimore County.



Mel'čuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N., and Mantha, S. (1988).

*Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques*. Presses de l'Univ. de Montréal.



Petitjean, S., Duchier, D., and Parmentier, Y. (2016).

XMG 2: Describing description languages.

In M. Amblard, P. de Groote, S. Pogodalla, and C. Retoré, eds., *Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996-2016) - 9th International Conference, LACL 2016, Nancy, France, December 5-7, 2016, Proceedings*, pp. 255–272.



Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014).

Extended phraseological information in a valence dictionary for NLP applications.

In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pp. 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

# Bibliography V



Savary, A. (2008).

Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2), 1–53.



Savary, A. (2009).

Multiflex: A Multilingual Finite-State Tool for Multi-Word Units.

In S. Maneth, ed., *Implementation and Application of Automata*, pp. 237–240. Springer Berlin Heidelberg.

preprint: <http://www.info.univ-tours.fr/~savary/English/papersASgb.html#CIAA09>.



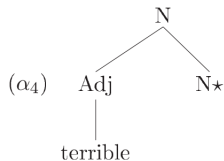
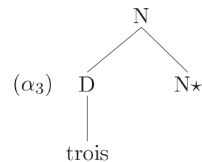
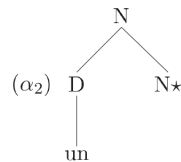
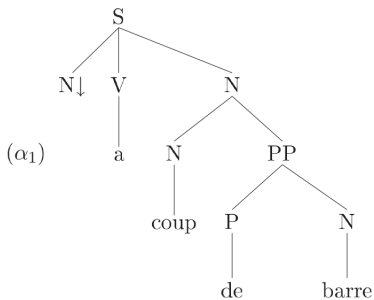
Tolone, E. and Sagot, B. (2011).

*Using Lexicon-Grammar tables for French verbs in a large-coverage parser*, p. 183–191.

Springer Verlag.

ISBN 978-3-642-20094-6.

# Insertion of adjuncts in LTAGs



# MWE long-distance dependencies in LTAGs

