Towards a French object-oriented MWE lexicon in XMG

Agata Savary [1,2], Simon Petitjean [2], Laura Kallmeyer [2], Timm Lichte [2]

[1]Université de Tours, France

[2]University of Düsseldorf, Germany

PARSEME-FR consortium meeting
15-16 January 2018
Marseille

Properties of MWEs

Types of properties

- Defective property – excludes a literal interpretation of the MWE, e.g.:
  - Defective agreement: ***grands-mères***
- Restrictive property – reduces the number of possible surface realizations of the MWE with respect to the literal reading, e.g.:
  - Restrictive lexical selection: ***retourner sa veste*** vs. *#retourer son blouson*
  - Restrictive agreement: *je **vide mon sac*** vs. *#je vide son sac*
  - Restrictive diathesis: ***les carottes sont cuites*** vs. *#on cuit les carottes*
  - Restrictive modification: *il **mène une vie** de riche* vs. *#il mène une vie*
  - Restrictive dependencies between determiners and modifiers: *j'**ai envie** de le faire*, *j'**ai** une **envie** folle de le faire*

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" $\succ$ "prohibt passive"
- "allow a possessive determiner"
  John pushed the/my door
  $\succ$ "impose a possessive determiner"
  John ***broke*** *his/our* ***fall*** 'John made his/our fall less forceful'
  $\succ$ "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

TAL-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)]

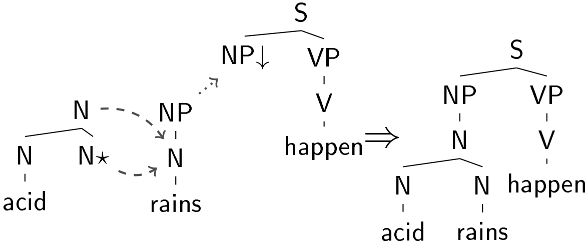
Recommendations

Requirements for a lexical encoding framework for MWEs [Lichte *et al.*(2016)]

- machine- and human-readable,
- representing specific **irregularities** of MWEs,
- friendly to **scale-wise** modeling,
- **factorized** (to avoid redundancies),
- **flexible** (to encode unforeseen properties),
- with a rigorous **denotational semantics** (to avoid vagueness and inconsistencies).
- easy to integrate in a **computational grammar**.

Tree Adjoining Grammars (TAGs)

- Elementary trees (ETs): initial tress (ITs) ∪ auxiliary trees (ATs)
- Tree rewriting: substitution & adjunction



MWEs in LTAGs

[Abeillé and Schabes(1989)]

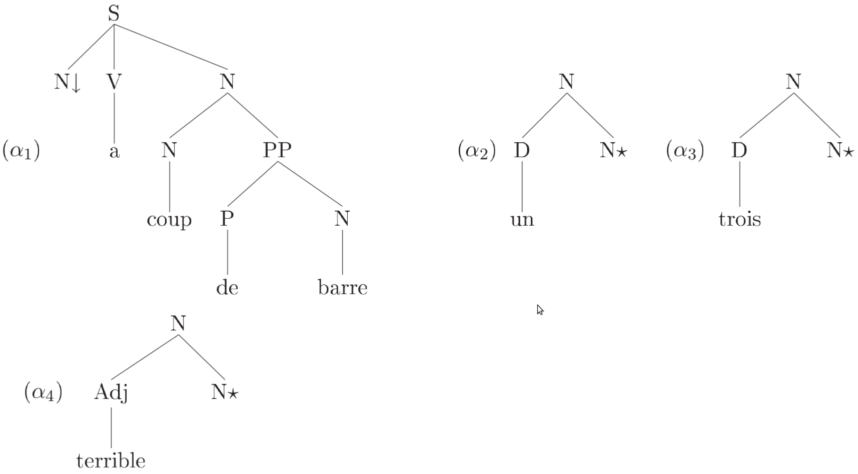
- Representing **discontinuities** (cf. *extended domain of locality*)
  - discontinuities in the internal structure of a MWE ⇒ visible in ETs, handled by **substitution**

    to take something with a pinch of salt
  - insertion of adjuncts ⇒ handled by **adjunction**

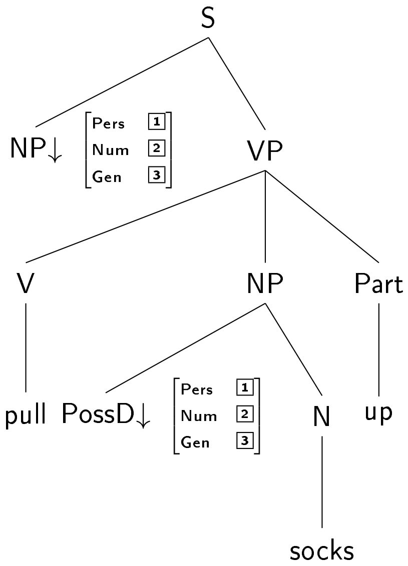
    a whole bunch of NP
- Dependencies between arguments at **different depths** in the ETs are naturally expressed

  She pulled her/#its socks up

Insertion of adjuncts in LTAGs



MWE long-distance dependencies in LTAGs



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 redudancy 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 uper ones)

XMG [Crabbé *et al.*(2013)]

- 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

FTAG – 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 FTAG with a (compositional) semantic dimension

Morphology

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

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

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

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

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

Trivial classes

propename $\rightarrow$ N$\diamond$

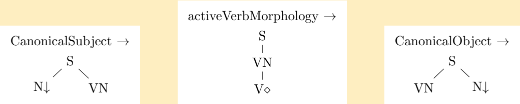
noun $\rightarrow$ N$\diamond$

CliticT $\rightarrow$ CL$\diamond$

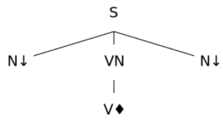
stddeterminer $\rightarrow$ N (D$\diamond$ N$*$)

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

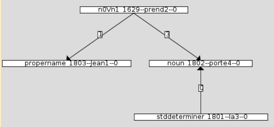
Metagrammar tree fragments inherited by n0Vn1



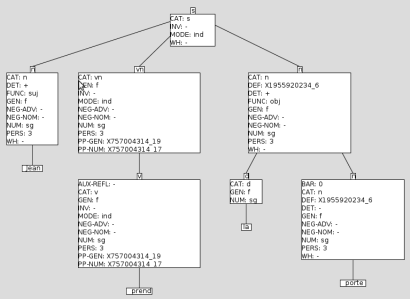
Grammar tree



Derivation tree

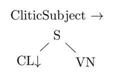


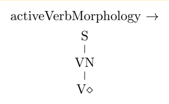
Derived tree

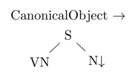


From metagramar to parsing: **n0Vn1** (*Il prend la porte*)

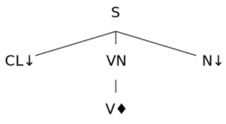
Metagrammar tree fragments inherited by n0Vn1



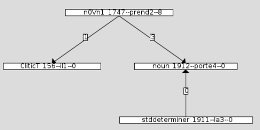




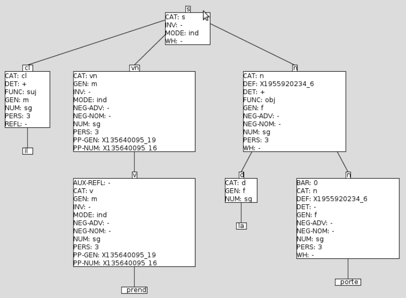
Grammar tree



Derivation tree

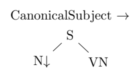


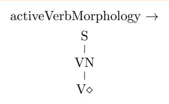
Derived tree

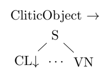


From metagramar to parsing: **n0Vn1** (*Jean la prend*)

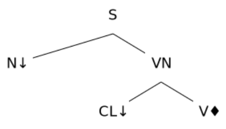
Metagrammar tree fragments inherited by n0Vn1



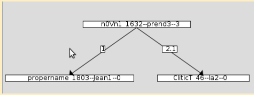




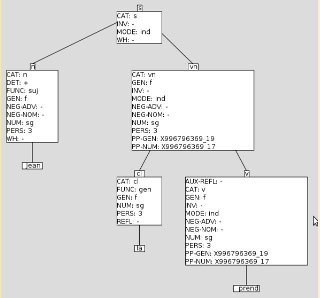
Grammar tree



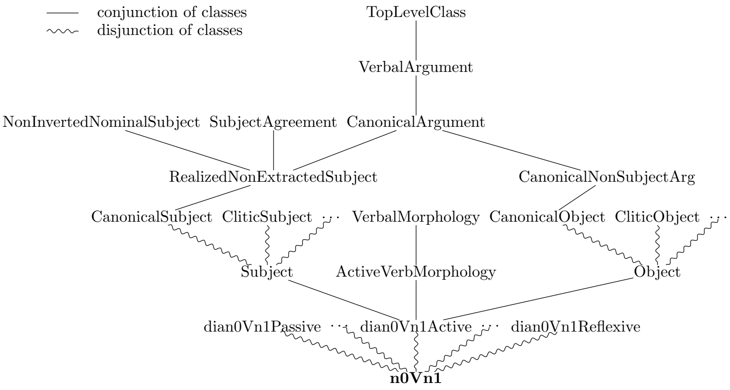
Derivation tree



Derived tree



Class hierarchy



MWEs have more or less regular properties

prendre la porte

- (More) **regular** features:
  - The subject is free and agrees with the verb:
    *Jean/il/elle **prend la porte***
    *Jean, que nous ne voulons pas ici, **prend la porte***
  - The verb inflects freely: ***Prend la porte**!*
- (More) **idiosyncratic** features:
  - The object is lexicalized: *#Jean prend la sortie*
  - The object cannot be:
    cliticised: *#Jean la prend*
    extracted: *#La porte que Jean prend*
    modified: *#Jean prend la grande porte*
  - The verb cannot be passivized: *#La porte est prise par Jean*

Adding MWEs to the metagrammar

Strategy 1 (applied here)

- **reuse** existing tree fragments for the (more) regular properties
- **duplicate** and **modify** existing tree fragments for slightly irregular properties
- **create** new tree fragments for (more) idiosyncratic properties

Strategy 2 (todo)

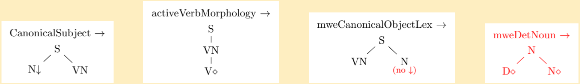
- add **features** to the MWE **lexical entries** marking the non allowed properties
- add **features** with opposite values to existing tree fragments to exclude parses if the features from the lexicon and from the tree do not unify
- Risk: this implies modifying the initial metagrammar

MWE lemmas with co-anchors

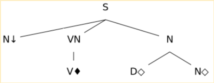
```
class mweLemmePrendreLaPorte
{
  <lemma> {
    entry <- "prendre";
    cat   <- v;
    fam   <-    mwen0VDetNActive;
    coanchor ObjDetNode -> "la"/d;
    coanchor ObjNode -> "porte"/n
  }
}
```

From metagramar to parsing: **mwen0VDetNActive** (*Jean prend la porte*)

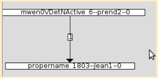
Tree fragments inherited by mwen0VDetNActive



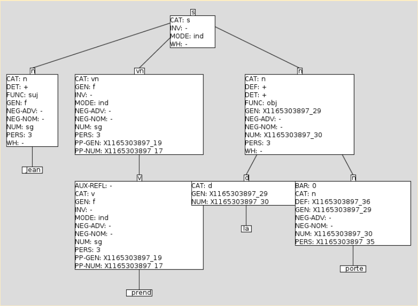
Grammar tree



Derivation tree

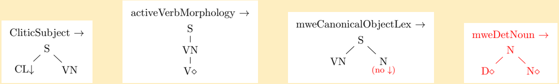


Derived tree

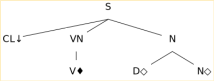


From metagramar to parsing: **mwen0VDetNActive** (*Il prend la porte*)

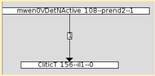
Tree fragments inherited by mwen0VDetNActive



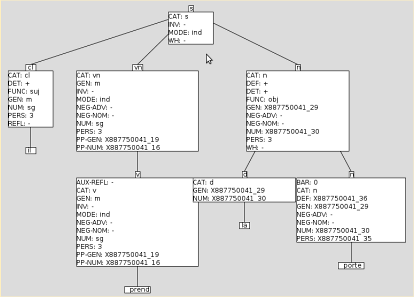
Grammar tree



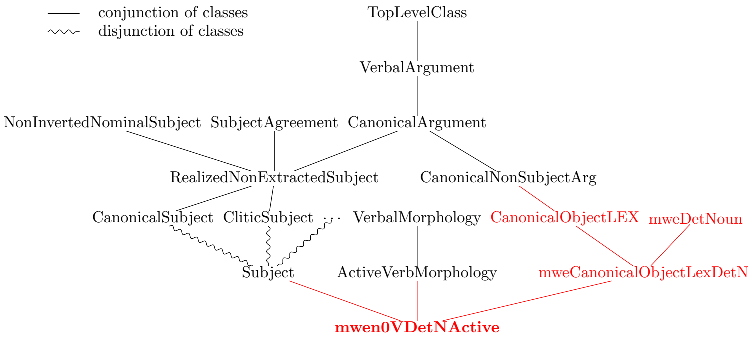
Derivation tree



Derived tree

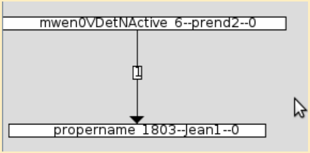


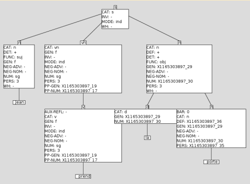
Modified class hierarchy



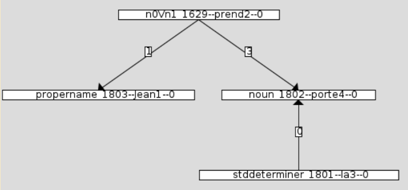
Two readings: *Jean prend la porte*

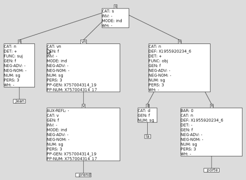
Idiomatic reading





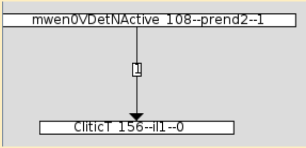
Compositional reading

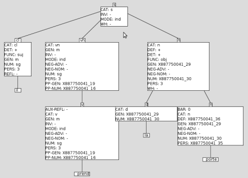




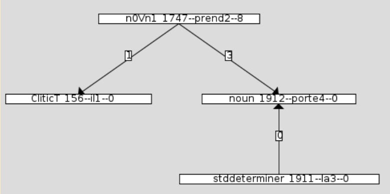
Two readings: *Il prend la porte*

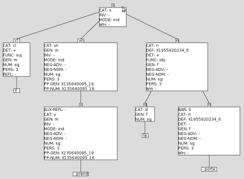
Idiomatic reading





Compositional reading

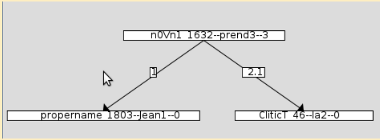


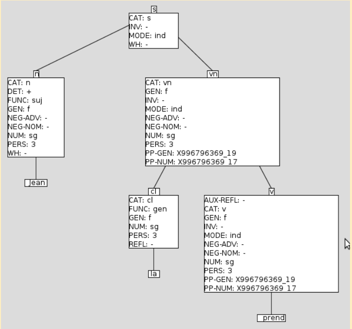


One reading: *Jean la prend*

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 modelled 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
- Handle **morphological features** in lexicon co-ancors
- Implement the **feature-based strategy**
- 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 Superieure*.

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 Coference 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.

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.

Savary, A. (2008).
Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches.
Linguistic Issues in Language Technology, 1(2), 1–53.

Bibliography V

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.