LVF-lemon – Towards a Linked Data Representation of "Les Verbes français"

Ingrid Falk, Achim Stein
Institut für Linguistik/Romanistik, Universität Stuttgart
first.second@ling.uni-stuttgart.de

Abstract

In this study we elaborate a road map for the conversion of a traditional lexical syntactico-semantic resource for French into a linguistic linked open data (LLOD) model. Our approach uses current best-practices and the analyses of earlier similar undertakings (*lemonUBY* and *PDEV-lemon*) to tease out the most appropriate representation for our resource.

Keywords: linguistic linked open data, LLOD, French valence lexicon, lemon, pdev, uby

1. Introduction

In this paper we investigate how a traditional lexical syntactico-semantic resource, namely "Les Verbes Français" (*The French Verbs*, henceforth abbreviated LVF) (Dubois and Dubois-Charlier, 1997; François et al., 2007), can best be converted into a standardised and normalised linked open data model. Our motivation is twofold. On the one hand we aim to explain and make more accessible the encoded linguistic knowledge. On the other we plan to make LVF more interoperable with and comparable to other linguistic resources, as for example corpus data and subcategorisation lexicons.

To convert LVF to LLOD format the following steps have to be undertaken:

1. RDF conversion
2. Data modelling: the content needs to be modelled in terms of well established vocabularies. These may be general vocabularies like RDFS, OWL, SKOS, linguistic vocabularies (lemon, LexInfo, OLiA, LMF), or finally vocabularies specific to LVF.
3. Linking the data.

While the first step is rather straightforward in our case, the second and third step require substantially greater research effort due to the traditional way this resource has been designed and developed.

A major strength of LVF lies in its syntactic and semantic description. We will therefore primarily investigate how this information and the intricate syntax-semantics interface can best be represented within a linked data framework.

2. The LVF resource

"Les Verbes Français" (Dubois and Dubois-Charlier, 1997; François et al., 2007) is a detailed and extensive lexical resource providing a systematic description of the morpho-syntactic and syntactico-semantic properties of French verbs. The basic lexical unit are readings of the verbs, determined by their acceptable syntactic and semantic context.

The LVF covers roughly 12 300 verbs with a total of 25 610 usages (readings). Each reading is associated with an elaborate morphologic, syntactic and semantic description.

In the following we use sample entries for the verb *élargir* (enlarge) to illustrate LVF's basic layout and give an idea of the underlying lexicographic and representation choices, in particular with respect to the syntactic and semantic description.

LVF lists four readings for *élargir*, illustrated by the sample usages shown in Table 1.

Semantic description. Each reading is characterised by a semi-formal semantic description, called *opérateur*, which is meant to represent its meaning, e.g. `r/d+qt large` for **élargir 01** (cf. Table 1a). The meaning of **élargir 01** is thus associated with the semantic primitive `r/d` ('render/become' indicating transformation). This way the readings were assigned one of a finite set of prototypical semantic predicates which, in a subsequent step, allowed their grouping into 14 well established semantic classes (*Transformation* or *Change of State* for **élargir 01**). Accordingly all readings are associated with two semantic descriptions: the *opérateur* and the *semantic class*.

Syntactic constructions. Each reading is coupled with a schematic representation of its acceptable syntactic constructions. These *schemes* encode the syntactic arguments (and some adjuncts) and indicate possible semantic realisations e.g. whether the subject (object) may be animate or plural, whether a syntactic argument refers to a manner or location. Some of the syntactic constructions assigned to **élargir 01** and **02** and the information they encode are shown in Table 1b.

In addition the lexicon provides inflectional information and indicates whether and how adjectival and nominal derivation is possible.

As this example shows, in LVF the syntactic and semantic descriptions are closely linked. The methodology of the elaboration of the semantic classes is explained in detail by François (2008). A more general introduction to LVF is provided by François et al. (2007).

3. The Linguistic Model: Lemon

The most prominent standard (meta-)model for building Linked Data lexicons and dictionaries is *lemon* – The Lexicon Model for Ontologies, (McCrae et al., 2012). Its main purpose is to link lexical linguistic data with the structured semantic information shared via the semantic web.

More specifically it was designed to meet the following challenges:

id	example[a]	*opérateur*	sem. primitive	sem. class
01	On *élargit* une route.	`r/d+qt large`	render/become	Transformation
02	Cette veste *élargit* Paul aux épaules.	`d large`	become+adj.	Transformation
03	On *élargit* ses connaissances.	`r/d large abs`	render/become, figurative	Transformation
04	On *élargit* le débat à la polititque étrangère.	`f.ire abs VRS`	caused directional move, figurative	Enter/Leave

(a) The four readings illustrated by sample sentences and their semantic description.

[a]Literal translations – 01: One broadens a road. 02: This jacket expands Paul's shoulders. 03: One widens one's knowledge. 04: One extends the debate to foreign policy.

id	schema	encoded information
01	A30	intransitive with adjunct, inanimated subject
	T1308	transitive, human subject, inanimated direct object, instrument adjunct
	P3008	pronominal, inanimated subject, instrument adjunct
02	N1i	intransitive, animated subject, prep. object w. prep. *de* (*of*)
	A90	intransitive with adjunct, subject human or thing
	T3900	transitive, inanimated subject, object human or thing

(b) Syntactic descriptions

Table 1: LVF entries for *élargir*

- Separation of the lexicon and knowledge (ontology) layers
- Linguistically sound structure based on LMF
- Linking to data categories, in order to allow for arbitrarily complex linguistic description. In particular this facilitates integration with annotated corpora
- RDF-native form to enable leverage of existing Semantic Web technologies

In a nutshell the *lemon* model consists of a list of *Lexical Entries* representing the words. These entities are on one the hand related to ontology entities providing a linguistic description. On the other hand they are connected via a *sense* property and a *LexicalSense* entity to ontology entities which represent their meaning.
The base component of the model is the *lemon core* which allows to represent a simple lexicon. Additional modules are proposed for modelling more sophisticated aspects of lexical representation:

- Linguistic Description
- Variation
- Phrase structure
- Syntax and mapping
- Morphology

Since in this contribution we are particularly interested in syntax and the syntax-semantics mapping, we are mainly concerned with the **Syntax and mapping** module which models syntactic frames and their mapping to logical predicates in a knowledge base (ontology).

Syntax and syntax-semantics interface. *Lemon* models grammatical relations and categories and in particular subcategorisation frames based on the LexInfo ontology: Figure 1 examplifies the mapping of syntactic and semantic arguments according to the *lemon* model.

4. Which Lemon Model for LVF?

From our description in the previous sections it has become apparent that there is no obvious way to convert the LVF lexical representation to the Lemon (meta-)model. The main difficulty stems from the syntactic and semantic description which in LVF are intricately interleaved whereas Lemon aspires at a clear-cut distinction.
Based on this in the following we investigate how this problem is addressed for two other comparable lexical resources, namely UBY (Eckle-Kohler et al., 2012) and PDEV (Maarouf et al., 2014).

4.1. lemonUBY

UBY (Eckle-Kohler et al., 2012) is a network of interlinked lexical semantic resources. It is similar to the *lemon* lexicon model in that it is based on LMF and externally defined data categories from ISOcat. The differences stem mainly from UBY's objective to fully cover a wide range of heterogeneous lexical resources.
The conversion to *lemonUBY* was achieved mainly by mapping the UBY LMF representation to the *lemon* LMF implementation.
The important point with respect to our task is the modelling of the **syntactic and semantic descriptions**. In *lemonUBY* subcategorisation frames are represented using LMF data categories (linking to ISOcat) instead of the LexInfo ontology used by *lemon*. Subcategorisation frames and senses are not explicitly linked but instead they are implicitly connected by the mapping of syntactic and semantic arguments. A syntactic frame is mapped to the set of thematic roles realising its syntactic arguments. Thus the syntax-semantics linking is made explicit, but the lexical meaning of the frame is presumed to be a composition of the meaning of the semantic arguments.
This conception is opposed to the position adopted in the PDEV project as we will see in the next section.

4.2. PDEV-lemon

PDEV, the Pattern Dictionary of English Verbs (Maarouf et al., 2014), is an empirically constructed resource where each

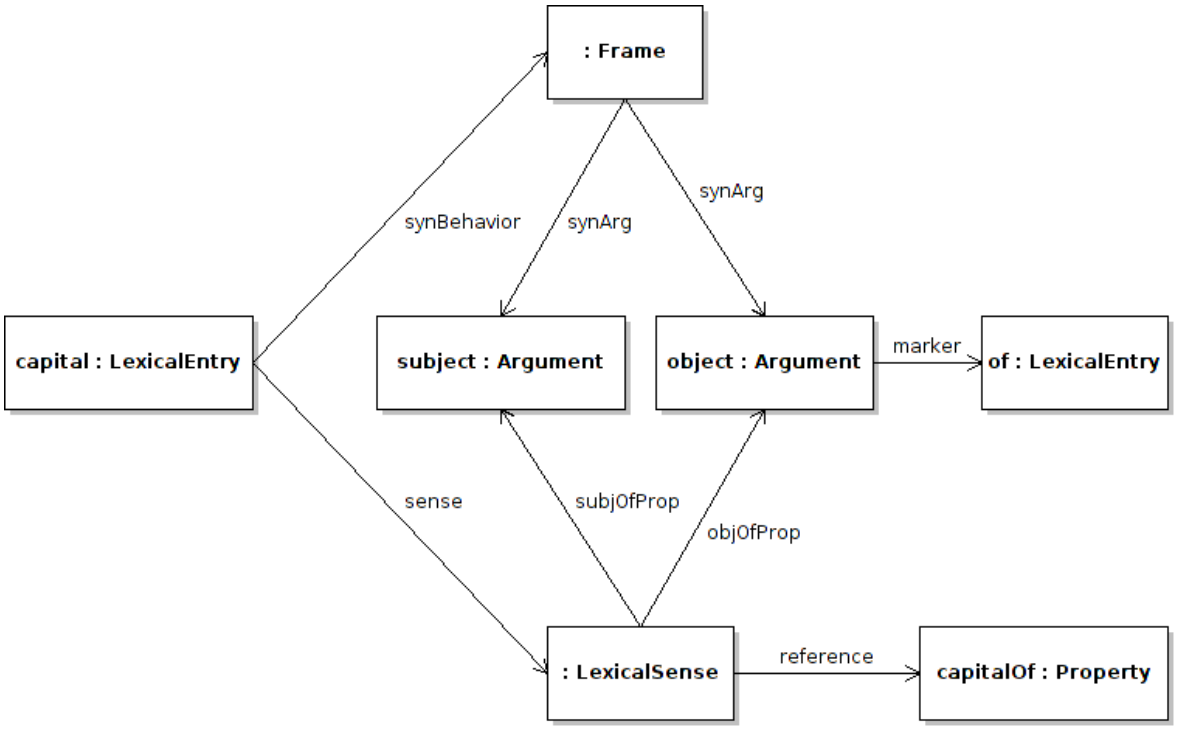

Figure 1: Syntax and syntax-semantics map in Lemon for *capital*.

verb is associated with its pattern of use. Patterns are attested in corpora and in essence represent a particular meaning.

In *PDEV-lemon* the lexical entries are the verbs. These are linked via `lemon:synBehavior` to syntactic patterns (or syntactic frames), modelled as subclasses of `lemon:Frame`. The structure of a pattern and the properties and categories of its arguments are represented using an extension of the LexInfo ontology.

The primary issue when creating *PDEV-lemon* was how to map a syntactic frame, selected by a lexical entry, to a meaning. Lemon does not provide direct links between a frame and a lexical sense. Lexical meaning of a frame is modelled either by instantiating it in the LexInfo ontology or through the mapping of syntactic and semantic arguments.

These solutions did not seem appropriate since, according to PDEV, syntax and semantics do not always map neatly onto each other (e.g. in the case of phraseological or idiomatic expressions). Therefore the two additional properties `:frameSense` and `:isFrameSenseOf` (in analogy to *lemon*'s `:sense` and `:isSenseOf`) were added to *PDEV-lemon* in order to link frames with lexical senses.

This reflects the view that the lexical meaning of the frame is not always related to the meaning of its arguments.

In addition *PDEV-lemon* provides an ontology of *semantic types* defining semantic properties as e.g. *animate*, *human*, etc.

5. The LVF-lemon model

Based on these investigations we propose the following road map for the creation and implementation of LVF-lemon:

LVF morphology can be appropriately represented by the *lemon* core model and its morphology module.

Syntactic constructions. LVF schemes can be easily mapped to *lemon or LMF* syntactic arguments and the *lemonUBY* ontology could at least partly be reused for the syntactic representation. However, in the LVF schema representation the syntactic arguments are further specified as follows.

First, they are assigned the following 9 semantic and/or syntactic types: *human*, *animal*, *thing*, *thing or phrase*, *phrase*, *human and plural*, *thing and plural*, *human or thing*. Since this representation does not correspond entirely but is related to semantic types defined for example in the PDEV ontologies, we propose to introduce LVF specific classes to represent this type system and to link them (via for example the `skos:related`) whenever possible to the related PDEV classes. Second, LVF has an inventory of 8 semantic roles which in some cases are used to further specify the syntactic arguments. The semantic roles consist of four types of location roles (representing the current location, an original location, a destination, and both an origin and a destination), a temporal role, a modality role (manner, measure, quantity), a cause role and an instrument or means role. As for the semantic types these are obviously related to well known role inventories (as in VerbNet or FrameNet) but there is no one-to-one correspondence. We therefore plan to represent these roles again as LVF classes and to relate them to the thematic roles of VerbNet, for example.

Semantic descriptions. For its semantic description LVF uses *semantic classes* and the *semantic primitives* in the *opérateur*. These need to be represented as LVF specific classes. Since both the *semantic classes* and *semantic primitives* are based on well established linguistic theories (cf. (Pinker, 1989; Jackendoff, 1983) among others), it should be possible to relate them to existing linked data resources along the lines of FrameNet and VerbNet.

Linking. A lexical entry in *LVF-lemon* is a verb reading, linked to *lemon:Frame* instances corresponding to the LVF readings. The syntactic component of the schemes can be represented by the *lemon:syntacticBehavior* relation. The link of the *lemon:Frame* instance with the corresponding semantic class can then be established via the lemon (or skos) *broader* property.

In the following, as a proof of concept and for illustration purposes, we describe a possible implementation of the road map laid out previously for the verb *élargir*, shown in Table 1 (focusing on syntax, semantics and syntax-semantics interface). In Figure 2 we present sample (pseudo) implementations in rdf turtle for the reading 01, using available *uby* 2a and *pdev* 2b vocabulary.

There would be four lemon entries for this verb, one for each of the four readings listed in the LVF (cf. Table 1) and each of these entries would correspond to a `lemon:LexicalSense`. However, the `lemon:LexicalSense` must remain underspecified, since LVF implements a different view of `LexicalSense` than it is understood by *lemon*[1]. More precisely, LVF does not give a traditional description of the lexical sense of the verb reading but rather the sense is suggested by semantic description(s) associated with these readings. Therefore, for a LLOD semantic representation of LVF entries we need a conceptual representation of the elements used in LVF for this semantic characterisation. We identified two such elements. The first is the semantic class (`Transformation` for the `élargir` readings, cf. Section 2.) and the second are the semantic primitives contained in the *opérateur* (in the case of *élargir* the main components of the latter are the primitives `r/d` – render/become, `d` – become, and `f.ire` – caused directional move). A possible conceptual representation would be for example LVF classes `lvf:SemanticClass` and `lvf:OperateurPrimitive`. The instances corresponding to the *élargir* readings could then be linked to the corresponding `lvf:SemanticClass` and `lvf:OperateurPrimitive` via the *lemon* or skos *broader* property. As already mentioned, the semantic class assigned to the *élargir* readings is called `Transformation` which is related (maybe distantly) with the VerbNet or FrameNet `Change of State` semantic classes. Ideally, these similarities would need to be explicited and fleshed out in order to obtain as much interoperability as possible.

In addition, the *opérateur* also contains other semantic cues, as for example the indication of a quantitative (`+qt`) and directional (`VRS`) meaning component, which could also be represented in a more normalised way.

Each of the lexical entries can be connected via the property `lemon:synBehavior` with `lemon:Frame` instances corresponding to the schemes used by LVF to describe the syntactic constructions selected by the verb in this particular reading. In the LVF vocabulary the following syntactic functions are encoded by the schemes: subject, object, prepositional complement and adjunct. While representations of the former are present in both the *uby* and *pdev* vocabularies, the representation of adjuncts needs to be fleshed out by using available Olia[2] or ISOcat[3] elements. Table 2 shows a possible more normalised representation of the schemes as subcategorisation frames. In this normalised form subcategorisation frames can be easily represented as linked open data using the *lemon* formalism, as shown for example in Eckle-Kohler et al. (2012).

A *lemon* subcategorisation frame consists of (representations of) syntactic arguments (`lemon:SyntArg`) which in turn are mapped to `lemon:SemArg` or thematic roles to represent the syntax-semantics interface. While the *lemonUby* resources represent this mapping, the LVF resource does not directly provide access to this type of information. The syntactic arguments in the LVF representation are nevertheless associated with semantic information of various kinds. It is for example specified if the subject or a complement are human, animals or things. For these semantic types, *PDEV-lemon* provides linked data representations which we could use. There is also, albeit not systematically, information referring to thematic roles. Thus it is for example indicated wether a complement or adjunct has a manner, temporal or instrumental role. To represent these it would be therefore necessary to define LVF specific semantic roles, which could be related for example with the semantic roles present in VerbNet.

6. Open Questions

As we saw in the previous sections, most of the LVF representation choices have related counterparts in existing LLOD repositories. However, while some can almost immediately be converted to LLOD, for others there is no straightforward way to achieve this. In particular, compared to modern conceptual frameworks used in LLOD representation, the LVF conceptual system is not completely specified and made explicit. Moreover, the LVF sometimes confuses classes which in current lexical ontologies are clearly kept apart.

An example is the way in which LVF specifies the syntactic arguments. Here it is necessary to clarify and better define the underlying concepts used for the definition of the syntactic/semantic types and the thematic roles. A further problematic issue is the encoding of prepositions in the schemes. Here the information provided by the LVF is valuable but not entirely systematic, and it is currently not clear how to best convert it to LLOD.

Further problematic issues are the semantic primitives and the semantic classes. While these can of course easily be represented as LVF classes, their linking to existing lexical LLOD elements is not straightforward. Nevertheless, such a linking would be beneficial for a comparison with other resources, and it should be feasible, since the LVF semantic concepts partly draw on the same fundamental linguistic theories as existing LLOD resources.

7. Conclusion

In this paper we discuss ways to convert a traditional French valence lexicon ("Les Verbes Français") into a linguistic

[1] The "*lemon*" way would be to associate the lexical entry (verb reading) with an ontology class.

[2] OLiA: The Ontology of Linguistic Annotation, `https://datahub.io/dataset/olia`

[3] ISOcat: a Data Category Registry, `http://www.isocat.org/`

```
@prefix rdf: <rdf uri> .
@prefix lemon: <lemon uri> .
@prefix ubyCat: <uby Olia categories> .
@prefix ubyVN:  <uby VerbNet vocabulary> .

:elargir-01 a  lemon:LexicalEntry ;
    ubyCat:partOfSpeech ubyCat:verb ;
    lemon:broader :Transformation [ a :
        SemanticClass ];
    lemon:broader :rendre-devenir [ a :
        OperateurPrimitive ];

    lemon:synBehavior [
         ## A30 intransitive frame with adjunct
            a lemon:Frame ;
            ubyCat:subject [
                 a lemon:Argument ;
                 ubyCat:syntacticCategory ubyCat:
                     nounPhrase ;
                 ## represent argument is "Thing"
                      ] ;
lemon:synBehavior [
         ## T1308 transitive with adjunct
            a lemon:Frame ;
            ubyCat:subject [
                 a lemon:Argument ;
                 ubyCat:syntacticCategory ubyCat:
                     nounPhrase ;
                 ## represent argument is "Human"
                       ] ;
            ubyCat:complement [
                 a lemon:Argument ;
                 ubyCat:syntacticCategory ubyCat:
                     nounPhrase ;
                 ## represent argument is "Thing"
                       ] ;
            :adjunct [
                 a :Adjunct ;
                 ubyCat:syntacticCategory ubyCat:
                     prepositionalPhrase ;
                       ] ;
                 ## represent adjunct is "Thing"
                     and "Plural"
                       ] ;

lemon:synBehavior [
         ## P3008 pronominal with adjunct
            a lemon:Frame ;
            ubyCat:subject [
                 a lemon:Argument ;
                 ubyCat:syntacticCategory ubyCat:
                     nounPhrase ;
                 ## represent argument is "Thing"
                 ## represent pronominal marker "
                     se"
                 ] ;
            :adjunct [
                 a :Adjunct ;
                 ubyCat:syntacticCategory ubyCat:
                     prepositionalPhrase ;
                       ] ;
                 ## represent adjunct is "Thing"  ]
                     and "Plural"
                  ] ;
]                  .
:Transformation rdfs:seeAlso ubyVN:Other_cos-45-4
     .
```

(a) Using *uby* vocabulary.

```
@prefix rdf: <rdf uri> .
@prefix rdfs: <rdfs uri> .
@prefix lemon: <lemon uri> .
@prefix pdev: <pdevlemon uri> .

:elargir-01 a  lemon:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:verb ;
  lemon:broader :Transformation [ a :
      SemanticClass ];
  lemon:broader :rendre-devenir [ a :
      OperateurPrimitive ];

  lemon:synBehavior [ a lemon:Frame ;
    ## A30 intransitive frame with adjunct
    pdev:subject [ a lemon:Argument ;
      pdev:syntacticCategory pdev:NounPhrase ;
        pdev:SemanticType [
          a pdev:PdevSemanticType ;
          rdfs:label "Physical Object"@en ; ] ;
                        ] ;
   lemon:synBehavior [ a lemon:Frame ;
    ## T1308 transitive with adjunct
    pdev:subject [ a lemon:Argument ;
      pdev:syntacticCategory pdev:NounPhrase ;
      pdev:SemanticType [ a pdev:
          PdevSemanticType ;
        rdfs:label "Human"@en ; ] ;
                ] ;
    pdev:directObject [ a lemon:Argument ;
      pdev:syntacticCategory pdev:NounPhrase ;
      pdev:SemanticType [ a pdev:
          PdevSemanticType ;
        rdfs:label "Physical Object"@en ; ] ;
                        ] ;
    :adjunct [ a :Adjunct ;
      pdev:syntacticCategory pdev:
          PrepositionalPhrase ;
      pdev:SemanticType [ a pdev:
          PdevSemanticType ;
        rdfs:label "Physical Object"@en ; ] ;
            ] ;
    ## represent adjunct is "Plural"
                  ] ;
   lemon:synBehavior [ a lemon:Frame ;
    ## P3008 pronominal with adjunct
    pdev:subject [ a lemon:Argument ;
      pdev:syntacticCategory pdev:NounPhrase ;
      pdev:SemanticType [ a pdev:
          PdevSemanticType ;
        rdfs:label "Physical Object"@en ; ] ;
      ## represent pronominal marker "se"
                ] ;
    :adjunct [ a :Adjunct ;
      pdev:syntacticCategory pdev:
          PrepositionalPhrase ;
      pdev:SemanticType [ a pdev:
          PdevSemanticType ;
      rdfs:label "Physical Object"@en ; ] ;
      ## represent adjunct is "Plural"
            ] ;
                  ] ;
               .
```

(b) Using *pdev* vocabulary.

Figure 2: Example rdf (turtle) listing for reading 01 of *élargir*, using *uby* respectively *pdev* vocabulary.

Schema	Name	Example	Normalised Representation
T$abcd$	transitive	T1308	subject,directObject,adjunct
		T3900	subject,directObject
P$abcd$	pronominal	P3008	subject,directObject,adjunct,reflexive marker
Aab	intransitive	A30	subject,adjunct
Nab	intransitive	N1i	subject,deObject

Table 2: Mapping of schemes to possible representations as subcategorisation frames. We used ISOcat identifiers wherever possible.

linked open data representation. The benefits are twofold. First, the underlying lexicographic and linguistic principles are translated into a modern vocabulary and are made explicit. Second it allows integration and interoperability with other linguistic resources, in particular corpus data. Our analysis of existing meta-models, namely *lemon*, and use cases (*lemonUBY* and *PDEV-lemon*) disclosed means and ways to achieve this goal.

8. Bibliographical References

Dubois, J. and Dubois-Charlier, F. (1997). *Les Verbes français*. Larousse.

Eckle-Kohler, J., McCrae, J., and Chiarcos, C. (2012). lemonUby - A large, interlinked, syntactically-rich resource for ontologies. *SWJ (Semantic Web Journal)*. Dataset description in the Special Issue on Multingual Linked Open Data.

François, J., Le Pesant, D., and Leeman, D. (2007). Présentation de la classification des Verbes Français de Jean Dubois et Françoise Dubois-Charlier. *Langue française*, 153(1):3–19, March.

François, J. (2008). Entre événements et actions: les schèmes composés de constructions syntaxiques du dictionnaire 'Les verbes français' de J. Dubois et F. Dubois-Charlier. *LIDIL*, 37.

Jackendoff, R. (1983). *Semantics and Cognition*. MIT Press, Cambridge, Mass.

Maarouf, I. E., Bradbury, J., and Hanks, P. (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, Mass.