Gathering Valency Frames for Annotation and Batch Corrections

Mathilde Regnault
Institute of Linguistics
University of Stuttgart (Germany)
{author}@ling.uni-stuttgart.de

Abstract

Syntactic annotation is time- and resource-consuming, especially for historical and heterogeneous data. The Universal Dependencies (UD) framework provides a stable and cross-linguistically consistent annotation scheme, offering a crucial backbone for diachronic corpus studies. However, ensuring internal consistency within historical UD treebanks remains challenging due to syntactic variation and parser errors. We address this issue for Medieval and Early Modern French by integrating valency information into our corrections to support UD treebank maintenance. Valency frames were extracted from the *Profiterole* treebank (v. 2.7) and used to enrich *OFrLex* with structured valency information for Medieval French. Existing lexical resources such as *Lefff* are also exploited for Contemporary French. These valency frames are used to detect and correct inconsistencies in automatically annotated data through batch operations, thereby reinforcing UD guideline compliance and improving annotation coherence across diachronic stages. Preliminary experiments on Medieval French and exploratory annotation of Early Modern French data suggest that lexicon-informed error mining can reduce manual revision effort while strengthening the diachronic continuity enabled by the UD framework.

Keywords: valency, lexicon, error mining

1. Introduction

Extending and maintaining syntactically annotated corpora remains a time- and resource-intensive task, especially for historical language stages. In the Universal Dependencies (UD) framework, automatic pre-annotation with off-the-shelf parsers is common practice, but parser performance drops significantly when applied to non-contemporary varieties exhibiting freer word order and greater structural variation (Grobol et al., 2022).

This is the case for Medieval and Early Modern French (now MedFr and EMFr). Manual correction is necessary, but it can be supported by error-mining strategies targeting the argument structure. In this paper, we investigate how lexical valency information can guide corpus correction. We focus on the extension of the *Profiterole* treebank (Prévost and Stein (2013); Prévost et al. (2024), 9th-15th c.) and additional EMFr texts (Gabay et al. (2022), 16th-18th c.). We exploit two similar and LTAG-compatible lexicons: the *Lefff* for Contemporary French (Sagot, 2010) and *OFrLex* for MedFr (Sagot, 2019).

The contribution is twofold. First, we describe the automatic extraction of valency frames from the Profiterole treebank (v2.7) using GREW (Bonfante et al., 2018), in order to enrich *OFrLex* with corpus-attested argument structures. Second, we present ongoing work on using these valency frames to detect and correct systematic parsing errors in automatically annotated data through batch correction procedures.

By integrating lexical valency constraints into the UD correction workflow, we aim to bridge lexical resources and dependency treebanks in the annotation of historical French.

2. Related Work

Valency information plays a central role in argument attachment and predicate–argument structure modelling (Van den Eynde and Mertens, 2003). In lexicalized formalisms such as LTAG, a verb specifies the number of arguments it selects, their grammatical function, and their syntactic realization. In canonical configurations, up to three core arguments are distinguished: Arg0 (subject), Arg1 (direct object or subject complement), and Arg2 (second object or oblique), with optional locative extensions.

Within the Universal Dependencies (UD) framework, core arguments are represented through relations such as *nsubj, obj, ccomp*, and *xcomp*, while non-core dependents are typically encoded as *obl*. However, the annotation of *obl* dependencies does not always specify the status of argument or optional modifier. As a result, deriving valency frames from UD-annotated corpora requires additional criteria to separate argumental obliques from adjuncts and to reconstruct lexical subcategorization patterns.

Several strategies have been proposed to acquire valency lexicons. A first approach relies on existing linguistic resources and formal lexical descriptions (e.g. Gardent et al. (2006) for Contemporary French). For MedFr, however, no large-scale, machine-readable valency lexicon is available.

lire v87 100;Lemma;v;<**Suj**:cln|sn,**Obj**:(cla|qcompl|scompl|sinf|sn),**Objà**:(cld|à-sn)>;@CompInd, @CtrlSujObj,cat=v;%actif,**%passif**,%ppp_employé_comme_adj, %passif_impersonnel,%se_moyen_impersonnel

Figure 1: Example of a *Lefff* entry: the verb *lire* ('to read') accepts a nominal subject (mandatory), a nominal direct object, an infinitive or a completive as Arg1 and a nominal complement introduced by the preposition *à* as Arg2 ('to read sthg$_{Arg1}$ to sb$_{Arg2}$'). The entry contains additional information, for instance the possibility for passive voice.

A second approach consists of extracting subcategorization frames from corpora. Early work for English derived frames from raw corpora using statistical methods (Briscoe and Carroll, 1997; Carroll and Fang, 2004; Korhonen, 2002). Other studies rely on syntactically annotated corpora, for instance Zabokrtskỳ (2005) for Czech and Kupść and Abeillé (2008) for French. Their method links arguments to governing verbs, compacts frames by integrating clitic pronouns into their corresponding functions, and separates active and passive realizations.

Our work adopts a similar corpus-driven strategy but operates directly on dependency-annotated data in the UD framework. In our case, valency frames are extracted from the Profiterole treebank (v. 2.7), a UD-annotated corpus of MedFr, which provides manually revised syntactic analyses. This choice offers several advantages: cross-linguistic comparability, standardized relation labels, and compatibility with UD-based parsing tools. At the same time, it raises specific challenges, particularly the absence of an explicit argument–adjunct distinction and the underspecification of secondary dependencies. We therefore combine UD relations (*nsubj, obj, ccomp, xcomp, obl, expl, aux:pass*) with distributional and morphological heuristics to reconstruct valency frames compatible with the tradition of the *Lefff* (see fig. 1).

Linguistic research often relies on automatically parsed data. This raises an issue for historical languages because of the heterogeneous nature of the data and the lack of resources to obtain gold annotation. Therefore, pattern extraction requires an expert knowledge of both the language and the texts. Working on several resources can also be a challenge, especially with concurrent annotation frames (Stein, 2024).

By grounding valency extraction in UD annotation, our approach contributes to ongoing efforts within the UD community to develop language resources and tailor resources for linguistic studies.

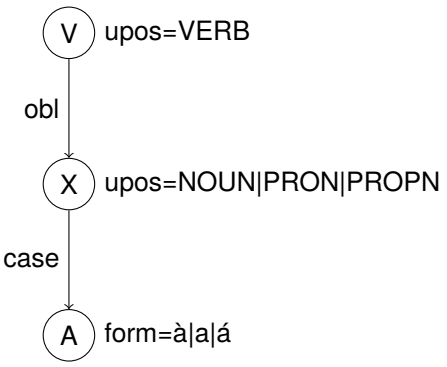

Figure 2: Example of a graph extracting candidates for the indirect object, as implemented in GREW

3. Methods

This section presents the methodology adopted to support UD treebank maintenance through lexicon-informed error mining. Our approach consists of two main steps: first, we automatically extract verb valency frames from a gold-standard UD treebank in order to construct a database of predicate–argument patterns and their syntactic realisations. Second, these frames are exploited to detect systematic mismatches between expected argument structures and automatic or converted annotation, enabling batch corrections.

The procedure combines automatic extraction using graph-pattern queries with targeted manual validation for structurally complex or diachronically unstable constructions. The use of UD-annotated data for valency extraction ensures the compatibility with UD guidelines while tailoring lexical resources to the specific syntactic properties of MedFr. The following subsection details the extraction phase.

3.1. Gathering Valency Frames

We used the gold-annotated texts of the SRCMF-UD Profiterole treebank (v2.7) to build a database of verb valency frames. Following Van den Eynde and Mertens's definition of valency grammar as a specific subtype of dependency grammar, UD dependency structures can be partially converted into *Lefff*-style argument structures. Argument extraction is performed using GREW. The graph representation makes it possible to jointly exploit syntactic dependencies and linear order information.

First, we gather the candidates for arguments, categorized in canonical valency slots (Arg0, Arg1, Arg2), and map UD relations to the syntactic functions listed in *OFrLex*. For instance, for the indirect object, we gather separately dative complements, nominal (see fig. 2) and infinitive complements introduced by the preposition *à*.

- **Arg0**: In a canonical TAG-based valency framework, Arg0 corresponds to the subject.

In MedFr, however, overt nominal or pronominal subjects are optional and often absent in early texts. We therefore exclude canonical nominal and pronominal subjects from automatic queries and treat them as default (though optional) realizations.
Clausal and verbal subjects are retained. Impersonal subject pronouns are identified through the *Cattex* label (its original tagset, Prévost et al. (2013)) *PROimp* (pronoun + impersonal) in XPOS and cross-checked with the UD relation *expl*. Impersonal subject pronoun *il* ('it') is not treated as a lexical paradigm element.

- **Arg1**: It covers
  - direct objects (*obj*),
  - finite clausal complements (*ccomp*),
  - infinitival complements (*xcomp* and *obj* headed by an infinitive),
  - subject predicates in copular constructions (*cop*).

  Different realizations (nominal, adjectival, prepositional complements, verbal or infinitival clauses) are encoded in the valency frame. Pronominal realizations are detected via *obj* dependencies restricted to clitic pronouns.
- **Arg2** (indirect object) is typically introduced by *à* (*Objà*) or *de* (*Objde*). It is possible to register two realisations for Arg2 (e.g. *paroler de... à...*, 'to speak about... to...'), provided that there is no Arg1. These complements are extracted via *obl* dependencies combined with prepositional information. However, the current version of *Profiterole* does not distinguish arguments from modifiers in obliques. To address this, we use a distributional heuristic: if a verb alternates between a prepositional complement (*à/de*) and the corresponding clitic pronoun (dative *li/lor*, genitive *en*), the complement is treated as a strong candidate for Arg2 status. This heuristic reduces noise but does not eliminate ambiguity. In particular:
  - *Objà* may be confused with locative complements,
  - *Objde* may overlap with source complements,
  - diachronic variation yields unstable prepositional marking,
  - some verbs allow multiple competing realisations.

  As noted by Amiot et al. (2020), verbal transitivity in the history of French is "rather chaotic". Consequently, certain frames required manual resolution. In the absence of a comprehensive reference lexicon for MedFr, automatic frame merging was not feasible. Non-argumental complements (*Loc, DLoc*, additional obliques) are stored separately and may host complements introduced by other prepositions (e.g. *avec* 'with', *sur* 'on'/'about').

Second, the number of hits per category is attributed to each lemma. For example, the verb *dire* ('to say' or 'to tell') registers 379 occurrences of direct objects, which indicates it is transitive. There are two candidates for an indirect object: nominal complements introduced by *à* (78) and *de* (43). However, the number of pronouns confirms this status for *à*-complements only, with 302 occurrences of dative clitic pronouns against 1 for *de*-complements.

Apart from core and optional arguments, other information has to be represented in the lexicon, for example the possibility for passive voice. Not all transitive verbs may be used in this manner, and the lexicon requires separate entries for such past participles. Therefore, passive constructions had to be detected via the UD dependency *aux:pass*. The reflexive pseudo-paradigm *se* is also extracted via UD dependencies (*expl*), excluding the impersonal pronoun *il*. It is encoded as part of the predicate rather than as an argument.

Highly polyfunctional verbs such as *faire* were also treated manually due to their structural complexity. Six entries were necessary to describe all the use cases, as in our reference, the *Lefff*, for example[1]:

(1) ex. canonical entry
Or **fai** ton mialz
now do-IMP your best
'Now do your best'
Yvain, Chrétien de Troyes, v. 4184 (12th c.)

(2) ex. light verb
Fai le **venir**
make-IMP he-ACC.SG come-INF
'Make him come'
Quatre Livres des rois, p. 31 (12th c.)

(3) ex. with an attribute and a reflexive pronoun
Se chaschun ne **se**
unless everyone NEG REFL.ACC
fait **humble** comme che enfant,
make-PRS humble like this child
'Unless everyone makes themselves humble like this child'
De la erudition, J. Daudin, p. 394 (14th c.)

[1] All examples of MedFr come from the BFM (Guillot et al., 2018).

This procedure (Regnault, 2022b) yielded an initial set of valency frames for 1,885 verbs (Regnault, 2022a) extracted from the Profiterole treebank, complemented by manually curated entries (see annex for detailed count). Automatic pruning of frames exceeding the maximum of three argument realisations can be considered. In our case, precision was necessary in order to parse as many sentences as possible with the corresponding TAG metagrammar (Regnault et al., 2019) and annotating them according to UD guidelines. In a second step, we aligned extracted lemmas with those of *OFrLex* to propagate attested argument structures across lemma variants, broadening the coverage of the lexicon.

3.2. Batch Corrections

The extracted valency frames are integrated into GREW-based correction workflows[2] targeting recurrent inconsistencies in the *Profiterole* treebank (Prévost et al., 2022), more specifically passive voice annotation and argument structure validation. Not all verbs marked by the passive voice received the proper *aux:pass* dependency in the *Profiterole*, making their extraction difficult for linguistic research. Only verbs compatible with passive formation should receive *aux:pass*, so we checked the 6,438 existing *aux* annotations and examined the transitive verbs employed with the *estre* auxiliary (when *avoir* is to be expected, which amounted to 436 cases) with a GREW rule. Missing verbal entries were added to the lexicon, and manual checks were necessary, especially for unaccusative verbs, allowing both 'be' and 'have' auxiliaries. This procedure resulted in the addition of 358 *aux:pass* dependencies. The *nsubj:pass* and *csubj:pass* dependencies have been automatically added, as well as *expl:pass*, but with a filter on the form, only allowing subject pronouns. The dependencies *obl:agent* however all necessitated a manual check, because they can be mistaken by general obliques, even when its head is an agentive noun.

Beyond passive normalization, we also addressed the annotation of indirect arguments. Many oblique complements did not receive a secondary label, in particular *arg* for indirect objects. Candidates for *obl:arg* were identified using a GREW lexical filter restricting the query to verbs selecting an Arg2. In order to avoid annotating modifiers as arguments, we added a lexical filter to the head of the complement, excluding nouns with a "time" feature from the lexicon and other lexical values taken from the corpus. This led to an increase of 464 *obl:arg* dependencies. These scripts will be useful for future expansions of the treebank.

To evaluate the portability of this workflow, we pre-annotated the EMFr *FreEM* corpus (Gabay et al., 2022) with the HOPS parser (Grobol and Crabbé, 2021) with the Flaubert model (Le et al., 2020), chosen for its proximity with Contemporary French.

First, we cross-check dependencies and lexicon entries. For example, in the parser-annotated version, there are 177 *nsubj:pass* dependencies for only 174 *aux:pass*. Some intransitive verbs also appear to be annotated with the passive voice, like *naître* ('to be born').

Second, we design graph-rewriting rules tailored to our data. Verbs which do not have the "passive" feature in the lexicon received a simple *aux* dependency (6 changes), and arguments were corrected according to the auxiliary's status (112 changes).

Although this stage remains ongoing, the rewriting scripts informed by *Lefff* valency frames can be applied to pre-annotated EMFr corpora as a support for maintenance or as an off-the-shelf correction workflow[3].

These interventions illustrate how valency-informed constraints can serve as a validation layer over UD dependency graphs, improving both internal consistency and cross-version reproducibility.

4. Conclusion

The maintenance and extension of historical UD treebanks remain effort-intensive tasks, particularly in the presence of diachronic variation and unstable argument structures. In this paper, we presented a corpus-driven extraction of valency frames from the gold-annotated *Profiterole* treebank, resulting in an initial inventory of 1,885 verbs enriched with structured argument information and aligned with *OFrLex*.

By extracting valency directly in UD dependency annotations, we provide a resource that is both theoretically interpretable and operational for treebank maintenance. The valency frames offer support to the detection of errors in argument structure and makes batch corrections possible and improves consistency with UD guidelines. While the large-scale exploitation of these frames for automatic revision remains ongoing, preliminary experiments indicate that lexicon-informed validation can substantially reduce manual intervention.

Beyond the case of Medieval French, this approach illustrates how UD-based lexical extraction can strengthen diachronic continuity and enhance reproducibility across treebank releases, especially when correction scripts are made openly available.

[2]These workflows are available in the "tools" section of the repository and may be linked to specific commits.

[3]This resource is available under: `github.com/mregnault/freem-treebank`.

5. Bibliographical References

Dany Amiot, Claire Badiou-Monferran, Bernard Combettes, Benjamin Fagard, Christiane Marchello-Nizia, and Maj-Britt Mosegaard Hansen. 2020. Catégories invariables. In Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, and Tobias Scheer, editors, *Grande Grammaire historique du français*, pages 856–961. de Gruyter.

Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*. Wiley-ISTE.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. *arXiv preprint*.

John Carroll and Alex C. Fang. 2004. The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *International Conference on Natural Language Processing (IJCNLP 2004)*, pages 646–654, Berlin, Heidelberg. Springer, Springer.

Claire Gardent, Bruno Guillaume, Guy Perrier, and Ingrid Falk. 2006. Extraction d'information de sous-catégorisationa partir du lexique-grammaire de Maurice Gross. *TALN 2006*.

Loïc Grobol, Sophie Prévost, and Benoît Crabbé. 2022. Is Old French tougher to parse? In *20th International Workshop on Treebanks and Linguistic Theories*, Sofia, Bulgaria.

Loïc Grobol and Benoît Crabbé. 2021. Analyse en dépendances du français avec des plongements contextualisés. In *Actes de la 28ème Conférence sur le Traitement Automatique des Langues Naturelles*.

Anna Korhonen. 2002. Subcategorization acquisition. Technical report, University of Cambridge, Computer Laboratory.

Anna Kupść and Anne Abeillé. 2008. Growing Treelex. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 28–39. Springer.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Sophie Prévost, Céline Guillot-Barbance, Alexei Lavrentiev, and Serge Heiden. 2013. Jeu d'étiquettes morphosyntaxiques CATTEX2009. Technical report, Technical report, École normale supérieure de Lyon.

Mathilde Regnault, Sophie Prévost, and Éric Villemonte de La Clergerie. 2019. Challenges of language change and variation: towards an extended treebank of medieval french. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 144–150.

Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Benoît Sagot. 2019. Développement d'un lexique morphologique et syntaxique de l'ancien français. In *26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, France.

Achim Stein. 2024. Réconcilier les formats d'annotation syntaxique pour faciliter l'analyse de la syntaxe française en diachronie. In *SHS Web of Conferences*, volume 191, page 11004. EDP Sciences.

Karel Van den Eynde and Piet Mertens. 2003. La valence : l'approche pronominale et son application au lexique verbal. *Journal of French language studies*, 13(1):63–104.

Zdenek Zabokrtskỳ. 2005. *Valency lexicon of Czech verbs*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague.

6. Language Resource References

Gabay, Simon and Clérice, Thibault and Gille Levenson, Matthias and Camps, Jean-Baptiste and Tanguy, Jean-Baptiste. 2022. *FreEM-corpora/FreEMlpm: FreEM LPM (Lemma, POS-tags, Morphology) corpus*. Zenodo.

Guillot, Céline and Heiden, Serge and Lavrentiev, Alexei. 2018. *Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique*. Presses de l'Université Paris-Sorbonne (PUPS), number 7 in Les états anciens des langues à l'heure du numérique.

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev, and Serge Heiden. 2024. Profiterole: un corpus morpho-syntaxique et syntaxique de français médiéval. *Corpus*, La constitution de corpus en diachronie longue. Méthodologies, objectifs et exploitations linguistiques et stylistiques.(25).

Prévost, Sophie and Grobol, Loïc and Dehouck, Mathieu and Regnault, Mathilde. 2022. *Corpus Profiterole*. GitLab.

Prévost, Sophie and Stein, Achim. 2013. *Syntactic Reference Corpus of Medieval French (SRCMF)*. ENS de Lyon/ILR Stuttgart.

Regnault, Mathilde. 2022a. *OFrLex-dev*. GitLab Inria.

Regnault, Mathilde. 2022b. *Valency finder*. GitHub.