

# Exploring predicate-arguments structures in texts to relate biological entities

Elisabeth Godbert<sup>1</sup>, Jean Royauté<sup>1</sup>,

Laboratoire d'Informatique Fondamentale de Marseille (LIF)  
CNRS UMR 6166 - Université de la Méditerranée  
Parc Scientifique et Technologique de Luminy, case 901  
13288 Marseille Cedex 9  
`{godbert,royaute@lif.univ-mrs.fr}`

**Abstract** : Verbal and nominal predicate structures present interesting linguistic properties. In biological texts they express relations between biological entities. We describe PredXtract, a platform which extracts predicate argument structures, and present the results.

**Mots-clés** : Predicate structure, Nominalizations of verbs, Dependency grammar, Link Parser.

## 1 Introduction

This paper focuses on the parsing of verbal and nominal predicate structures, expressed in a great variety of ways (Meyers et al., 2004). Defining a uniform representation for these structures is decisive to converge on a VerbNet or FrameNet representation (Wattarujekrit et al., 2004; Miyao et al., 2006) and for applications in information extraction (Meyers et al., 1998) and acquisition of semantic relations.

In this perspective, we have developed a robust platform, PredXtract, based on the Link Parser (Sleator and Temperley, 1991). This platform extracts verbal and nominal predicate argument structures (PAS) in texts and allows to exhibit relations between biological entities.

## 2 Classification of the predicates

Verbs and their nominalizations are the most productive predicates and have the same argument relations, where arguments play precise conceptual roles: subjects and complements, which are core arguments, and adjuncts. We present here seven important classes of predicates, defined from their core arguments.  
(i) Verbs accepting a direct object are grouped together in Class 1 and 2; in the corresponding predicate noun phrases (PNPs), the preposition *of* marks the

direct object. (ii) In Class 3 to 5, verbs do not accept a direct object, and in the PNP the preposition *of* marks the subject. (iii) Class 6 and 7 concern symmetric predicates with interchangeable arguments.

**Class 1:**  $N_0 V N_1 = N^{pred} \text{ of } N_1 \text{ by } N_0$ . Example: *IFN-gamma activates protein kinase C delta / activation of protein kinase C delta by IFN-gamma*.

**Class 2:**  $N_0 V N_1 \text{ Prep } N_2 = N^{pred} \text{ of } N_1 \text{ Prep } N_2 \text{ by } N_0$ . Example: *N<sub>0</sub> attributes a protein fragment to a sequence / attribution of a protein fragment to a sequence by N<sub>0</sub>*.

**Class 3:**  $N_0 V = N^{pred} \text{ of } N_0$ . Example: *the femoral head necroses / necrosis of the femoral head*.

**Class 4:**  $N_0 V \text{ Prep } N_1 = N^{pred} \text{ of } N_0 \text{ Prep } N_1$ . Example: *tryptophans fluctuates in gramicidin / fluctuation of tryptophans in gramicidin*.

**Class 5:**  $N_0 V \text{ Prep } N_1 \text{ Prep } N_2 = N^{pred} \text{ of } N_0 \text{ Prep } N_1 \text{ Prep } N_2$ . Example: *temperature decreases from 200 K to 70 K / decrease of temperature from 200 K to 70 K*.

**Class 6:**  $N_a V \text{ with } N_b = N^{pred} \text{ of } N_a \text{ with } N_b = N^{pred} \text{ of/between } N_a \text{ and } N_b$ . Examples: *genes interact with proteins ; interaction of genes with proteins / interaction of/between genes and proteins*.

**Class 7:**  $N_0 V N_a \text{ Prep } N_b = N^{pred} \text{ of } N_a \text{ with/to } N_b \text{ by } N_0 = N^{pred} \text{ of/between } N_a \text{ and } N_b \text{ by } N_0$ . Examples: *N<sub>0</sub> connects a new sequence with/to a cluster ; connection of a new sequence with/to a cluster / connection of/between a new sequence and a cluster*.

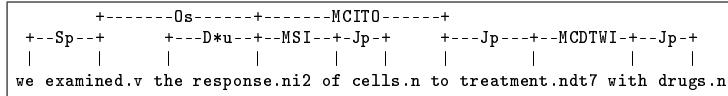
### 3 PredXtract, an extracting platform

PredXtract uses the Link Parser and its native Link Grammar (LG), a variant of dependency grammars (Sleator and Temperley, 1991). In LG, generic links attach verbs (M<sub>v</sub>p link) or nouns (M<sub>p</sub> link) to any preposition which introduces an NP. In order to mark, in the parses, the precise role of each argument of the predicates, we have defined specific argument links, which are searched during the extraction process. PredXtract is constructed with the following components.

**An extended lexicon.** Following Szolovits (2003), we have added in the grammar all the words of the "Specialist Lexicon" (SL), which includes UMLS terms (<http://www.nlm.nih.gov/research/umls>). We have then added a lexicon of genes and proteins extracted from corpus.

**A special PNP grammar.** According to our classification of the nominalizations, we have developed a grammar with 89 subclasses, where each subclass corresponds to a syntactic pattern with core arguments (including clauses with *that*) and adjuncts. Each nominalization belongs to one or more subclass.

The frame below gives an example of a short sentence with two nominalizations. The first one (*response*), concerns a subclass (noted ni2) of Class 4, where the preposition *to*, inherited from the verb, introduces the complement. The MSI link marks the subject introduced by the preposition *of*, while the MCITO link marks the complement introduced by *to*. The second shows the prepositional



use of *treatment*, here not saturated, corresponding to a subclass (noted **ndt7**) of Class 2, where the preposition *with* (link **MCDTWI**), inherited from the verb, introduces the complement.

**A verb-noun alignment module.** Rather than writing a grammar for verbs, which would have been very complex, we have defined a module that aligns verb arguments to nominalization arguments during a post-processing step. This module performs several tasks: (i) distinguish complements from adjuncts of verbs, by using the data of SL, and substitute the generic **MVp** link with a specific argument link when appropriate; (ii) identify each "verbal sequence", compound with a verb and a set of possible auxiliaries, negation, and modal verbs; (iii) identify arguments in passive or active voice, and interchangeable arguments.

**A recognition module of predicate structures.** For each parse of a sentence, all the predicates and their arguments are identified (argument links point on the heads of core arguments). Then the surface structure of each argument is reconstructed via the links, by using linguistic criteria. The reconstructed arguments can be NPs (most cases), clauses or adverbs.

**A filtering module of parses.** For each sentence, the parses (often several thousands) are re-ordered by attributing to each parse a score defined from several criteria. Among the main criteria: (i) a higher score is given to parses whose number of argument links is maximum in the case of multiple prepositional attachments to verbs or nouns; (ii) a specific score is calculated in the case of PNPs containing several nominalizations, to favour prepositional arguments attached to the head of the PNP.

We present below an example: from the sentence *Hyperoxic exposure induced an S-phase arrest associated with acute inhibition of Cdk2 activity and DNA synthesis*, 9168 parses were found and PredXtract outputs :

---

```

Nominalization 1: exposure
Nominalization 2: arrest
  subject or object: S-phase
Nominalization 3: inhibition
  direct object: Cdk2 activity
  direct object: DNA synthesis
Nominalization 4: synthesis
  subject or object: DNA
Verb 1: induced (verbal sequence: induced ; active)
  subject: hyperoxic exposure
  direct object: an S-phase arrest associated with acute inhibition of [...] synthesis
Verb 2: associated (verbal sequence: associated ; passive)
  direct object A: an S-phase arrest
  direct object B: acute inhibition of Cdk2 activity and DNA synthesis

```

---

This example shows a short sentence with six predicate structures. We can notice that (i) *exposure* has no argument, (ii) *inhibition* has two coordinated objects,

(iii) the role of the argument of *arrest* and *synthesis* is underspecified (subject or object), and (iv) the verb *associated* has two interchangeable arguments (object A and object B).

## 4 Results and conclusion

PredXtract has been evaluated with a corpus of 335 Medline abstracts, given by biology researchers. From the 3,500 sentences of this corpus, we have selected 700 random sentences; 300 of them have been used to finalize our system, and the evaluation has been done on the 400 others. Nominalizations represent 42.3% of all predicates. Because of the possibility of wrong segmentation of arguments, we have calculated two values for recall, precision and F-measure, with: [Case 1] only the true and complete arguments, [Case 2] the true and complete arguments and the true but incomplete arguments. In Case 1, the F-measure score is 0.78, for nominalizations and 0.77 for verbs. In Case 2, it is 0.85 for nominalizations and 0.88 for verbs. Thus, we observe a very small difference between values for nominalizations and verbs.

As PredXtract is based on very large lexicons, it is possible to say that PredXtract is a platform which extensively recognizes PAS, independently of the predicate type. The following step will be the identification of types of terms in arguments, by using term resources (UMLS terms, genes, proteins, etc.).

## Acknowledgments

Many thanks to Christine Brun and Bernard Jacq of LGPD-CNRS, for having supplied us with their corpus of Medline abstracts tagged with gene nouns.

## References

- Meyers, A., Macleod, C., Yangarber, R., Grishman, R., Barrett, L., and Reeves., R. (1998). Using nomlex to produce nominalization patterns for information extraction. *Proceedings of the COLING-ACL '98 Workshop on Computational Treatment of Nominals, Montreal, Canada.*
- Meyers, A., Reeves, R., Macleod, C., Szekeley, R., Zielinska, V., Young, B., and Grishman., R. (2004). The cross-breeding of dictionaries. *In proceedings of LREC-2004, Lisbon, Portugal.*
- Miyao, Y., Tomoko, O., Katsuya, M., Yoshimasa, T., Kazuhiro, Y., Takashi, N., and Tsujii., J. (2006). Semantic retrieval for the accurate identification of relational concepts in massive text bases. *In Proceedings of the COLING-ACL, Australia, pages 1017–1024.*
- Sleator, D. and Temperley, D. (1991). Parsing English with a Link Grammar. *Carnegie Mellon University Computer Science technical report, CMU-CS-91-196, Carnegie Mellon University, USA.*
- Szolovits, P. (2003). Adding a medical lexicon to an english parser. *In Mark Musen, editor, Proceedings of the 2003 AMIA Annual Symposium, pages 639–643.*
- Wattarujeekrit, T., Shah, P. K., and Collier, N. (2004). Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5.