# Definition and representation of complex function words in dependency annotation schemes: regularity, endogeneity and parsing accuracy

*José Deulofeu, André Valli, Carlos Ramisch, Alexis Nasr*
*Aix Marseille Univ, CNRS, Université de Toulon, LIS, Marseille, France*
andre-valli@wanadoo.fr, jose.deulofeu@gmail.com, {FirstName.LastName}@lis-lab.fr

Complex function words can be defined as combinations of two or more lexemes that have the distribution of a simple function word, such as complex prepositions (e.g. *in front of*), conjunctions (e.g. *as well as*) and determiners (e.g. *a lot of*). The question that we propose to address is: how should one represent complex function words in dependency syntax, and in particular in treebanks? This is particularly relevant in the context where multilingual schemes, such as Universal Dependencies, emerge (Nivre et al 2016). The goals of our contribution are (1) to discuss the problem on theoretical grounds and its consequences for NLP and linguistic models (2) to propose and discuss possible annotation schemes (and guidelines).

**The problem.** Our starting point is a lexicon of complex function words in French, focusing on prepositions and subordinating conjunctions (Ramisch et al 2016). When selecting entries, we adopted the operational definition that complex function words are chains of items fulfilling the same syntactic function as individual items. In implementing the definition, we favored formal criteria against semantic ones. However, comparing our results to current annotations in treebanks, we noticed a significant number of discrepancies (De Smedt et al 2016). We will show that the source of the discrepancies relies mainly on disagreements either in the definition of what is a "regular syntactic structure" or in the inner representation of the proposed multiword items. Then, we will propose a procedure to choose the most adequate representation with regard to parsing accuracy.

**What is a regular syntactic structure?** Syntactic regularity and semantic opacity are often used as (implicit) criteria in guidelines for complex function words (Kahane et al 2018). We will justify our choice to favor analysing constructions as regular structures whenever possible. For example, sequences composed of preposition + *que* + finite clause are considered as regular syntactic structures, by extending the valency set of prepositions to finite clauses. Consequently, we disregard sequences such as *pour que* (*for that*) as complex conjunctions. By contrast, we will explain why we did not extend this analysis to the adverb + finite clause sequences such as *bien que* (in spite of*)* and *alors que* (whereas), which are analysed as complex conjunctions.

**Representing the inner structure of complex functional words.** Because of their potential ambiguity, we want to represent the inner structure of complex function words (Candito & Constant, 2014). Figure 1 shows two sentences: (a) contains a literal reading of *à partir de* (lit. *to leave of*) meaning 'about leaving', whereas (b-e) contain the complex function word meaning 'after'. We want to distinguish both because, for automatic processing, they should not be represented in the same manner (e.g. they need to be translated differently into English). However, there are several possibilities to distinguish them in the syntactic annotation layer. The first 2 possibilities (b-c) use standard dependencies whereas the last 2 possibilities (d-e) use UD-style dependencies. We will discuss the advantages and inconvenients of each proposed annotation scheme, focusing on the issue of the choice between a **motivated** versus **arbitrary** structure of complex functional word. We will argue in favor of an endogeneous structure of functional complex words. That is, instead of choosing arbitrarily an item as the target of incoming dependency arcs (e.g. the first token), we propose that this target should be the item that can fulfill alone, in other contexts, the function of the whole expression. In examples (c) to (e) the target could be either the first or the last item, as both are prepositions. But the existence of complex prepositions introduced by noun phrases instead of prepositional phrases, like *histoire de* (lit. *history of*, in order to) and *le temps de* (lit. *the time of*, just after), argues in favor of choosing the last item as the target.

**On experimental or practical grounds: choosing a representation based in parsing results.** We will propose a way to evaluate the different representations by comparing the accuracy of a parser trained on different representations. It is hard to answer this question in general because the answer depends on the underlying parsing model. Nonetheless, we propose to conduct experiments using a transition-based dependency parser. This model makes local decisions, that affect the way syntactic trees are constructed. Given a treebank where the target complex function words of our lexicon are annotated, we propose representing them using the proposed structures shown in Figure 1 and train and evaluate a parser on such variants. We expect that the analysis of parsing performance and of the automatically constructed parsing trees will show whether some of these representations are more suitable to parsing.

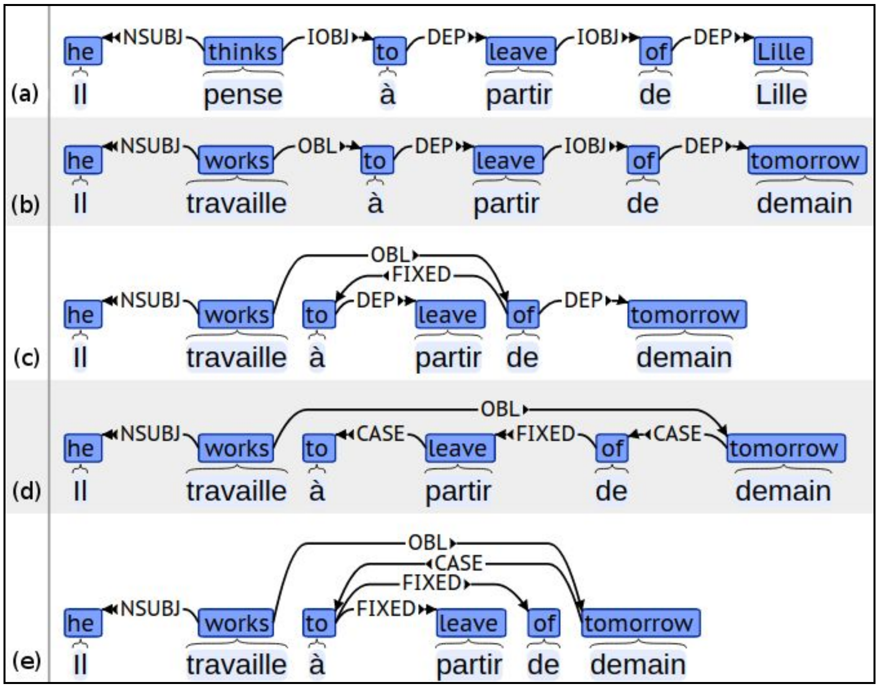
(a)
he NSUBJ thinks IOBJ to DEP leave IOBJ of DEP Lille
Il pense à partir de Lille
(b)
he NSUBJ works OBL to DEP leave IOBJ of DEP tomorrow
Il travaille à partir de demain
(c)
OBL
FIXED
he NSUBJ works to DEP leave of DEP tomorrow
Il travaille à partir de demain
(d)
OBL
he NSUBJ works to CASE leave FIXED of CASE tomorrow
Il travaille à partir de demain
(e)
OBL
CASE
FIXED
he NSUBJ works to FIXED leave of tomorrow
Il travaille à partir de demain

**Figure 1**: Five example sentences with identical POS sequences: (a) contains a literal occurrence of the sequence of words *à partir de* (about leaving)*,* whereas (b-e) show four representation proposals for the complex function word *à partir de* (after).

## References

Marie Candito and Matthieu Constant. 2014. *Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing*. In ACL 14 - The 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, USA. ACL.

Koenraad De Smedt, Victoria Rosén and Paul Meurer. 2016. *Studying consistency in UD treebanks with INESS-Search*. In Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14), Warsaw, Poland.

Sylvain Kahane, Kim Gerdes and Marine Courtin. 2018. *Multi-word annotation in syntactic treebanks: Propositions for Universal Dependencies.* In 16th international conference on Treebanks and Linguistic Theories (TLT), Prague, Czech Republic.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. *Universal dependencies v1: A multilingual treebank collection*. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Portorož, Slovenia.

Carlos Ramisch, Alexis Nasr, André Valli, José Deulofeu. 2016. *DeQue: A Lexicon of Complex Prepositions and Conjunctions in French*. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.