Multi-word annotation in syntactic treebanks

Propositions for Universal Dependencies

Sylvain Kahane
Université Paris Nanterre
Modyco (CNRS)
sylvain@kahane.fr

Marine Courtin, Kim Gerdes
Sorbonne Nouvelle
ILPGA, LPP (CNRS)
marine.courtin@etud.sorbonne-nouvelle.fr, kim@gerdes.fr

Abstract

This paper discusses how to analyze syntactically irregular expressions in a syntactic treebank. We distinguish such Multi-Word Expressions (MWEs) from comparable non-compositional expressions, i.e. idioms. A solution is proposed in the framework of Universal Dependencies (UD). We further discuss the case of functional MWEs, which are particularly problematic in UD.

1 Introduction

In every linguistic annotation project, the delimitation of lower and upper boundaries of the annotation units constitutes a basic challenge. In syntactic annotation, the lower boundaries are between morphology and syntax, the upper boundaries between syntax and discourse organization. This paper discusses the lower boundaries in syntactic treebank development. We place our analysis in the Universal Dependency framework (UD), which constitutes a large community of more than 100 teams around the globe (Nivre et al. 2016).

In this paper, we want to discuss the problem caused by idioms in syntactic annotation. The literature on idioms and MWEs is immense (Fillmore et al. 1988, Mel'čuk 1998, Sag et al. 2002, etc.). Our goal is not to mark the extension of MWEs on top of the syntactic annotation (see Savary et al. 2017 for a recent proposition). Our purpose is to tackle the impact of idiomaticity on the syntactic annotation itself. Most idioms (such as *kick the bucket* or *green card*) do not cause any trouble for the syntactic annotation because their internal syntactic structure is absolutely transparent (and it is precisely because they have an internal syntax that they are idioms and not words). Some expressions, however, such as *not to mention, heaven knows who, by and large, Rio de la Plata* (in English), are problematic for a syntactic annotation, because they do not perfectly respect the syntactic rules of free expressions.

We propose two contributions:

- For a coherent annotation it is crucial to distinguish **syntactically irregular** structures from **semantically non-compositional** units. These notions are highly correlated but distinct and we propose criteria to distinguish them.
- We explore different ways of annotating these two kinds of Multi-Word Expressions and their combinations in a syntactic treebank, with a special focus on functional MWEs.

Section 2 proposes a simple typology of MWEs opposing semantic compositionality and syntactic regularity. In section 3, we lay the basis of our analysis by discussing the syntactic units of a dependency annotation and point to problems in the current UD scheme (version 2.1). In section 4, we propose to analyze MWEs with an internal syntactic structure according to their level of syntactic regularity. We show how an MWE can be introduced into the current CoNLL-U format as a unit with its own POS. In section 5, we introduce two convertible dependency schemes for functional MWEs before concluding in section 6 with an example combining the MWE as a separated unit with the new convertible scheme for functional MWEs.

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 181–189,
Prague, Czech Republic, January 23–24, 2018. Distributed under a CC-BY 4.0 licence. 2017.

Keywords: dependency, annotation, syntax, UD, MWE, idioms

2 Idioms and syntactic irregularity

We distinguish idiomatic expressions from syntactically irregular constructions. Idiomaticity is a semantic notion and semantics has to be annotated apart from syntax.

Even if it is not our purpose to define idiomaticity here, let us give some thoughts to the matter. Following Fillmore 1988 (with his *encoding* and *decoding idioms*) or Mel'cuk 1998 (with his *phraseme* and *collocation*), we distinguish two levels of non-compositionality. We adopt the point of view of *encoding*: "Compositionality […] is to be distinguished from analysability, which pertains instead to the extent to which speakers are cognizant […] of the contribution that individual component structures make to the composite whole." (cf. Langacker 1987:457). An MWE is an *idiom* (i.e. *non-compositional*) if its components cannot be chosen individually by the speaker (*kick the bucket* is chosen as a whole and there is no possible commutation on its components).[1] An MWE is a *collocation* (i.e *semi-compositional*) if one of its component is chosen freely (the basis) and the other one (the collocate) is chosen according to the basis (in *wide awake*, *wide* can be suppressed and *awake* keeps the same contribution: *awake* is the basis and *wide* is a collocate expressing intensification with *awake*).

We also consider three levels of syntactic irregularity. First, natural languages contain some syntactic subsystems which do not follow the general properties of syntactic relations. For instance, most languages have particular constructions for named entities such as dates or titles. English has a regular construction N N, where the second noun is the head (*pizza boy, Victoria Lake*) but it also has a subsystem where the first noun is the head, used for named entities (*Lake Michigan*, *Mount Rushmore, Fort Alamo*). These subsystems are in some sense "regular irregularities", that is, productive unusual constructions. Similarly, English produces a high number of multi-word adverbs from a preposition and a bare noun as in *on top (of)* or *in case (of)*, thus forming another sub-system that does not conform to the typical syntactic system of English.

Second, languages have non-productive irregular constructions. Most of these irregular constructions are idioms, but some are compositional. This is the case of Fr. *peser lourd* 'weigh a lot/be significant', lit. weigh heavy, where *lourd* is an adjective that commutes only with NPs (*peser une tonne* 'weigh one ton').[2] Even the commutation with its antonym *léger* 'light' is impossible. Another example is Fr. *cucul la praline* 'very silly', lit. silly the praline. It is a collocation: the adjective *cucul* can be used alone and the NP *la praline* is an intensifier. The POSs of the units are clear, and the dependency structure can be reconstructed, but it is unusual to have an NP modifying an adjective.

We consider four cases of non-productive irregular constructions.

a. Structures with a clear POS and dependency structure but that function as a whole differently than their syntactic head: the coordinating conjunction headed by a verb *not to mention* (*they gave us their knowledge, not to mention their helpfulness*), the adjective *top of the range*, headed by a noun (as in *a very top of the range restaurant*), the French pronoun *Dieu sait quoi* 'heaven knows what', headed by a verb.

b. For some sequences, the POS are clear, but the dependency structure has to be reconstructed diachronically (the Fr. pronoun *n'importe quoi* 'anything', lit. no matter what)[3] or inversely, the dependency structure is clear but the POS have to be reconstructed (the adverb *by and large* – *by* being originally an adverb).

c. Other sequences have no clear internal dependency structure at all, while the POS remain clear: *each other*, Fr. *à qui mieux mieux* 'each trying to do better than the other', lit. to whom better better.

[1] An idiom can be semantically transparent (Svensson 2008). For example, it is quite clear that a *washing machine* is a machine that is used to wash something, but is an idiom because it is arbitrary that this denotes a machine for washing clothes and not a dishwasher or a high-pressure water cleaner. An idiom can even be semantically analyzable, cf. Gibbs 1994:278: "Idioms like *pop the question* [...], s*pill the beans*, and *lay down the law* are 'decomposable', because each component obviously contributes to the overall figurative interpretation."

[2] How the relation between *peser* and *lourd* must be analyzed in UD is not quite clear. *Lourd* should probably be analyzed as an `xcomp` of *peser* but if we do that we lose the fact that *lourd* is in the paradigm of NPs analyzed as `obj`.

[3] Diachronically, *quoi* is the subject of *importe* but now it is recognized as an object due to its position.

d. Some sequences have neither clear POS nor an internal structure in the language of the corpus: the adjective *ad hoc*, the proper noun *Al Qaeda*, and the Fr. SCONJ *parce que* ‘because’.[4]

	Compositional	Semi-compositional	Non-compositional
Regular construction	Typical syntax (*the dog slept*)	*[wide] awake, [heavy] smoker, rain [cats and dogs]*	*kick the bucket, green card, cats and dogs, in the light (of),* Fr. *pomme de terre* ‘potato’
Sub-system	Dates: *5th of July, tomorrow morning* Titles: *Miss Smith*	*Ludwig van Beethoven* in German (*van* is a Dutch word similar to Ger. *von*)	*on top (of), in case (of),* Fr. *à côté (de)* ‘next (to)’ Meaningful dates: *September 11th, 4th of July* *Mount Rushmore, Fort Alamo*
Irregular construction	Fr. *peser lourd* ‘weigh a lot’, lit. weigh heavy	Fr. *cucul la praline* ‘very silly’, lit. silly the praline	a) *not to mention, a lot (ADJ-er), top of the range,* Fr. *Dieu sait quoi* ‘heaven knows what’ b) Fr. *n'importe quoi* ‘anything’, *by and large* c) *each other*, Fr. *à qui mieux mieux* ‘each trying to do better than the other’, lit. to whom better better d) *ad hoc*, *Al Qaeda,* Fr. *parce que* ‘because’

Table 1. Different types of MWEs

Table 1 opposes degrees of syntactic regularity in the rows and semantic compositionality in the columns. In section 4, we will propose an annotation scheme for irregular constructions and for some non-compositional sub-systems.

3 MWE in UD

3.1 MWE and tokenisation

The tokenization of UD follows the underlying principle that tokens must be words or parts of words. A priori no token contains spaces (except well delimited cases of polysyllabic words) and therefore multi-word expressions are described syntactically and not morphologically. This is a vital choice for practical and theoretical reasons: Ambiguous sequences cannot be disambiguated on a morphological level without taking into account the whole sentence. Therefore, the alternative choice of multi-word tokens containing spaces is problematic: In the manual annotation process, creating the tokenization and the syntactic analysis at the same time is time-consuming, annotating a special link for MWE is much more user-friendly. For automatic parsing, too, a tokenization as a separate task that precedes the actual dependency annotation is redundant because both tools need a global view on the sentence – and syntactic parsers are specialized tools to do just that. Moreover, two annotations of the same sentence are harder to compare if they are based on different tokenizations and a spelling-based annotation makes that possible because it does not depend on the possibly ambiguous syntactic annotation itself.

Inversely, grouping Multi-Word Expressions together in a syntactic annotation scheme can at its most simple form always be achieved by introducing into the set of relations special ad hoc links for multi-words. UD makes use of this approach with the links `fixed`and `flat`[5] where no internal structure is annotated. In UD terms we could reformulate the purpose of the paper simply as: When must the `fixed` relation be used?

3.2 Problems with the MWE encoding in UD

This work springs from a recognition that the treatment of functional MWEs in UD is unsatisfactory for at least four reasons:

[4] Historically *parce* is the preposition *par* ‘through’ and the pronoun *ce* ‘that’, but this is not visible in today's orthography. The attribution of a POS to *parce* seems arbitrary and the French UD treebanks are subsequently incoherent: Fr-Original calls *parce* an ADV, Fr-Sequoia an SCONJ, and Fr-ParTUT has both versions.

[5] `flat` is a relation used for headless constructions (such as *Bill Clinton* for which is it not easy to decide which word is the head). This relation concerns productive and regular sub-systems and will not be discussed here.

1) The relation `fixed` is commonly used for MWEs with a very clear internal syntactic structure (see Figure 1).[6]

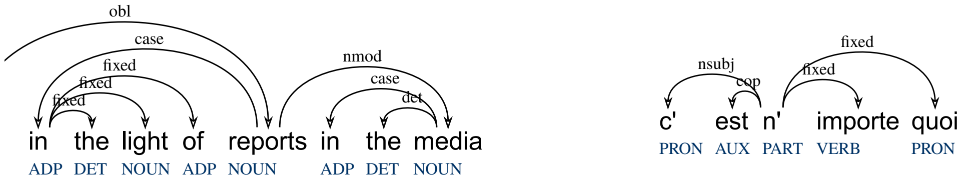


Figure 1. Analyses with `fixed` in En-PartTUT and Fr-Original

When analyzing them as `fixed` MWEs, we flatten the structure, losing precious information in the process, which will give us fewer instances of these syntactic relations on which to train our parser (cf. Gerdes & Kahane 2016's principles as well as the principles given on the UD introduction page). Moreover, the analysis is somewhat contradictory: If we recognize the POSs of the components (such as the verbal nature of *importe* in Fr. *n'importe quoi* 'anything', lit. no matter what), then we could also recognize the dependency relations that the tokens entertain.

2) Currently, the criteria to decide which constructions enter the realm of MWEs are insufficient and we observe a lot of discrepancies between different treebanks and even inside a single treebank.

For instance *along with* appears with three analyses. In En-ParTUT *along* is considered as the `case` marker of the noun phrase and *with* as *along*'s `fixed` dependent. On the other hand, En-Original mainly favors a compositional analysis with both *along* and *with* as `case` markers, but there is also one occurrence where *along* is a `cc` dependent of the noun phrase and *with along*'s `fixed` dependent.

Tables 2 and 3 give an overview of the usage of the MWE-relations in the English and French UD treebanks. When comparing the highlighted lines in the English and the French tables, we observe that the usage that annotators make from the three MWE relations `compound`, `fixed`, and `flat` go beyond what can be expected as language and genre differences and rather seems to indicate that the annotators understood the relations differently. This is corroborated by the high inter-corpus variation, for French, too. The two French treebanks Fr-FTB and Fr-Sequoia, for example, do not use `compound` at all. The significant number of observed incoherences in these two languages suffices to show that the UD annotation guide for MWE relations clearly deserves an overhaul in order to achieve a higher inter-language, inter-corpus, and inter-annotator annotation.

3)The POS of an MWE as a whole does not appear explicitly.

The assumption made is that the MWE will have the same POS as its syntactic head but many examples show that this is not the case. For example *not to mention* is a coordinating conjunction, a useful information for a syntactic parser that cannot be retrieved from the POS of its units.

[6] UD's definition of `fixed` refers to Sag et al. (2002) who say: "Fixed expressions are fully lexicalized and undergo neither morphosyntactic variation (cf. **in shorter*) nor internal modification (cf. **in very short*). As such, a simple words-with-spaces representation is sufficient. If we were to adopt a compositional account of fixed expressions, we would have to introduce a lexical entry for "words" such as *hoc*, resulting in overgeneration and the idiomaticity problem (see above)." Let us remark that, first, limits on modification do not imply weird lexical entries, as the example *in short* shows itself – the two words being in the lexicon anyhow. Second, and most importantly, an MWE can have constraints on modification for a specific meaning while still remaining transparent for the speaker, not only diachronically: *in short*, for example, is identifiable as a prepositional phrase, even if *short* is originally an adjective. This leads to multiple but syntactically constrained internal modifications of MWEs, not only in puns and journalistic style, but more generally also in ordinary coordinations and elisions as we will see below. Note also that the current 2.0 En-Original corpus consistently annotates *in short* (3 occurrences) and *for short* (1 occurrence) as a compositional prepositional phrase (`case-nmod`), contrarily to Sag's paper referenced in the annotation guide.

English	compound	fixed	flat
En-Original	**4,38 %**	**0,24 %**	**0,73 %**
En-Lines	**2,63 %**	**0,49 %**	**0,72 %**
En-ParTUT	**0,40 %**	**0,56 %**	**1,24 %**
total number of MWE	9194	966	1882
max freq variation between corpora	1107%	43%	59%
total nb links	11993	1091	2625
total frequency of links	3,58 %	0,33 %	0,66 %
total nb MWE types	7067	122	1215
average nb of occurrences per type of MWE	1,3	7,9	1,5
non-contiguous types	292	4	0

Table 2. Measures for MWE of the English UD v2

French	compound	fixed	flat
Fr-Original	**0,21 %**	**1,04 %**	**1,79 %**
Fr-FTB	**0,00 %**	**8,75 %**	**0,70 %**
Fr-ParTUT	**0,23 %**	**1,04 %**	**0,44 %**
Fr-Sequoia	**0,00 %**	**2,56 %**	**1,25 %**
total number of MWE	786	33190	9444
max freq variation between corpora	N/A	843%	411%
total nb links	877	55975	11858
total frequency of links	0,08 %	5,36 %	1,14 %
total nb MWE types	660	8544	7329
average nb of occurrences per type of MWE	1,2	3,9	1,3
non-contiguous types	24	58	0

Table 3. Measures for MWE of the French UD v2

4) The span of MWEs in the current UD scheme is questionable in some cases, especially concerning governed prepositions, which are not separated from the MWE itself (cf. *of* in Figure 2, below).[7]

4 Propositions for the encoding of MWEs in UD

All regular constructions from Table 1, including idioms, should be analyzed internally because:

1. Such a tree is syntactically more informative than any type of flattened structure where readily available syntactic relations have been removed.
2. We can expect a higher inter-annotator agreement on the syntactic relations if the annotation of MWE is kept independent from syntax, because of the difficulty of defining and recognizing MWEs
3. Equally, we can expect better parsing results because we have more instances of every relation and unknown idioms can obtain a correct parse, too.

The same holds for all compositional and semi-compositional constructions. We even go as far as proposing to analyze non-productive irregular constructions in case a) and b) by regular syntactic relations, but for some MWEs, we need means of encoding the POS of the whole expression because its POS is not identical to its head's POS. We propose to use `fixed` only for parts of c) and d) where the regular syntax does not provide appropriate syntactic relations.

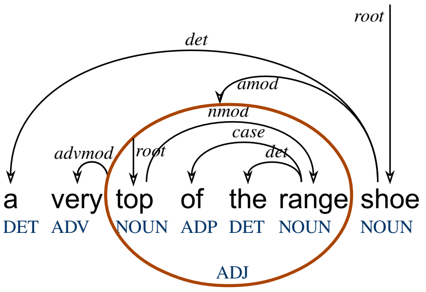
In some MWE of c) and d), some relations remain transparent and we could annotate partial structures whenever they are available. For example *à qui mieux mieux* contains a clear *à* <`case`- *qui* relation independent of the analysis of the rest of the expression.

[7] The preposition can be repeated (*According to the President and* ***to*** *the Secretary of State* – the repetition can disambiguate the scope of the shared element in the coordination) which seems incompatible with the `fixed` analysis favored in the English treebanks. In other languages, such as French, the repetition is quite systematic. In English, governed prepositions are particularly cohesive with their governor, giving us what is called *preposition stranding* in extraction (*the girl I talk to*). But even in this case, nobody denies that the verb *talk* subcategorizes a preposition phrase and that the preposition *to* is not part of the verb form. The fact that the preposition is not a part of the idiom becomes even clearer with expressions such as *in front of X*, where the subcategorized phrase can be suppressed (*she stopped in front*) or pronominalized (*in its front*). Note that the alternative classical dependency analysis where prepositional phrases are governed by prepositions results in a more coherent analysis because the governor (the verb or the expression) always forms a subtree with the sub-categorized preposition, independently of the extension given to the MWE.

For those remaining fixed relations, dependency distance measures would give more reliable result if the standard bouquet annotation (all words depending on the first token) would be replaced by a series of left-to-right relations connecting one word to its neighbor, because the absence of any recognizable syntactic relation rather implies some relation of simple juxtaposition than a structure headed by the first word.

The CoNLL-U format can easily be extended to allow for a fully expressive annotation of MWEs. One solution is to devote one specific column holding the idiomatic information (or equally, put this information into a specific attribute in the feature column of CoNNL-U). This choice does not allow embedding MWEs in one another. A better choice is to extend the current multi-word token format by adding a line for each MWE. This additional line could also include the POS of the whole expression.[8] It constitutes an additional unit that can constitute a node of a semantic graph. This could be combined with a specific MWE column or simply a specific feature in the additional line's FEATS column that distinguishes different types of non-compositionality, following the Parseme project: for instance idioms, light-verb constructions, and named entities.

In the following example, the governor of the MWE *top of the range* is *shoe*. But the head/root of the MWE is *top*.



1	a	a	DET	_	_	7	det
2	very	very	ADV	_	_	**3-6**	**advmod**
3-6	**top of the range**	_	**ADJ**	_	_	**7**	**amod**
3	top	top	NOUN	_	_	**3-6**	***root***
4	of	of	ADP	_	_	6	case
5	the	the	DET	_	_	6	det
6	range	range	NOUN	_	_	3	nmod
7	shoe	shoe	NOUN	_	_	0	root

Figure 2. UD analysis of the adjective *top of the range* (case a)Functional MWEs in UD

UD presents a particular problem with functional MWEs, because UD favors dependencies between content words (determiners and prepositions are dependents of the noun following them). It appears that the choice made by UD to have the prepositions as dependent of their complement is the source of some "catastrophes" (in the mathematical sense of the term) as soon as "prepositional" MWEs are in-volved (Gerdes & Kahane 2016). The goal of this section is to present the problem and to propose a solution to smooth it.

Let us consider the following examples illustrating what is often called a complex determiner (1a) and a complex preposition (1b):

1. (a) She asked me **a lot of** questions.
   (b) She lives **in front of** my house.

We can compare these sentences with (2a) and (2b):

2. (a) She asked me **many** questions.
   (b) She lives **near** my house.

According to the choices made by UD, we have dependencies between *asked* and *questions* in (2a) and between *lives* and *house* in (2b) (Figure 3)

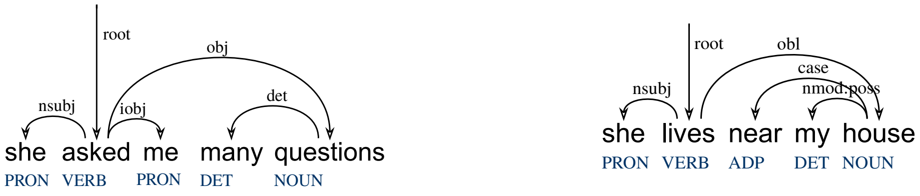


Figure 3. UD analysis of 2a and 2b

[8] Currently the format is only used for contiguous items. The format can be extended to non-contiguous expressions, e.g. we could have "3-5,7-8" as an index.

It is tempting to preserve these dependencies and to treat *a lot of* and *in front of* respectively as a complex determiner and a complex preposition. Let us first remark that *of* in these expressions is not part of the MWE, but is part of the sub-categorization of the MWE, by parallelism with verbal sub-categorization (cf. footnote 7, although the coherence of these expressions is higher and the preposition cannot always be repeated alone). In other words, the MWEs in question are *a lot* and *in front*. Theses MWEs are syntactically transparent and we do not want to analyze them with `fixed`. Two analyses are possible.

Analysis A respects the surface syntax and *of N* is treated as the complement (`nmod`) of the MWE. This is the most common analysis in the current English UD treebanks.[9]

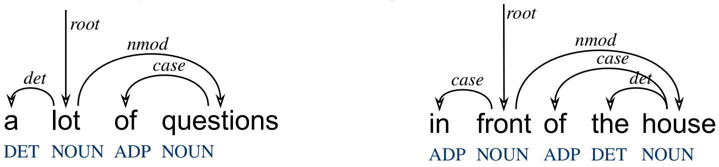


Figure 4. Analysis A for *a lot (of)* and *in front (of)*

Analysis B favors the relation between content words, as in the analyses of Figure 3. In this analysis, we propose to introduce special relations `det:complex` and `case:complex` when the dependents of `det` and `case` are MWEs.

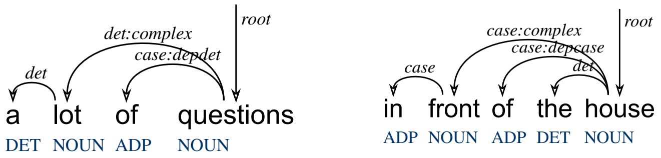


Figure 5. Analysis B for *a lot (of)* and *in front (of)*

The sub-categorized preposition *of* is governed by the complement noun. We introduce a feature on the case relation to indicate that this preposition is subcategorized by a dependent of the noun. We need to distinguish `case:depdet` and `case:depcase` because both can be present: *in front of a lot of houses*, where *front*, *lot* and the two *of* will depend on *houses*.

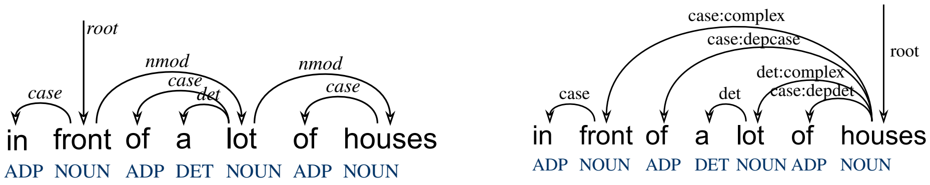

root
nmod
case
det
nmod
case
case
in front of a lot of houses
ADP NOUN ADP DET NOUN ADP NOUN
case:complex
case:depcase
det:complex
case:depdet
case
det
root
in front of a lot of houses
ADP NOUN ADP DET NOUN ADP NOUN

Figure 6. Analyses A and B for *in front of a lot of houses*

Both analyses A and B are interesting. It is possible not to choose and to allow the conversion from one analysis to the other. For that we need to enrich analysis A, by adding the subtype `:antidet` and `:anticase` to the standard `nmod` relations which go the other way in the B analysis (and are labeled `det:complex` and `case:complex`).

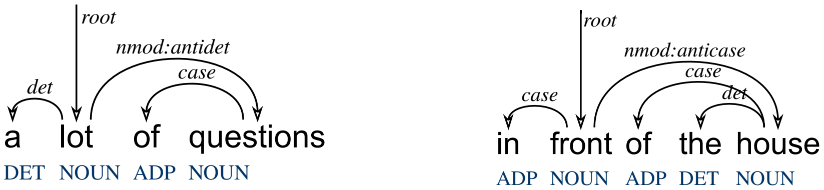
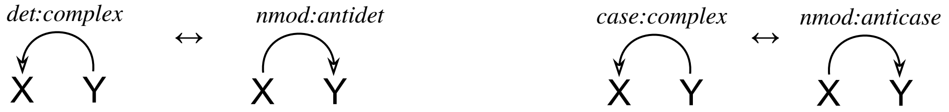


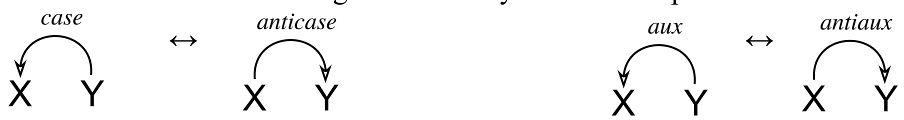
Figure 7. Enhanced analysis B for *a lot (of)* and *in front (of)*

[9] Since *quite a lot (of questions)* is possible, *a lot* has actually become an adverb (just like in *a lot better* – or other comparative adjectives) and the relation between *a lot* and the noun complement *of questions* should be of type `obl` and not `nmod` as it is in the current English UD treebanks. This irregular behavior of *a lot* can be captured by the introduction of an MWE unit as in Section 4.

Our rules of conversion are:



Similar rules could be used to get a surface syntax-based representation from UD:[10]



5 Conclusion

We have shown that irregular structures need to be introduced as units because we have to associate a POS to them. In cases a) and b) the internal structure is transparent but the POS of the complete unit is not predictable. In cases c) and d), where we use `fixed` relations, it is all the more necessary to indicate the POS of the MWE. For regular idioms, too, we can add the MWE as a unit.

For regular functional MWEs, we propose to add sub-types to the relation to capture the relations between content words, as well as the syntactic dominance relations. A tree does not allow expressing both types of relations at the same time, but the proposed sub-types relations can be converted from one to another.[11]

The two proposals are orthogonal and can be combined. For example, if we want to treat *a lot* as an adverb, we can have the analysis of Figure 8:

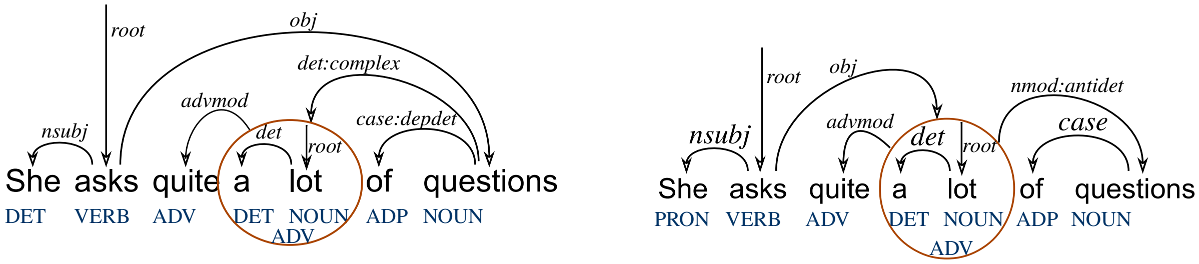

root
obj
det:complex
advmod
nsubj
det
root
case:depdet
She asks quite a lot of questions
DET VERB ADV DET NOUN ADP NOUN
ADV
root
obj
nmod:antidet
nsubj
advmod
det
root
case
She asks quite a lot of questions
PRON VERB ADV DET NOUN ADP NOUN
ADV

Figure 8. Analysis A and B for *quite a lot of questions*

The proposed schemes and distinctions clarify some underspecifications in the current UD scheme that lead to incoherent analyses. The usage of subtypes fits in unintrusively into the current scheme and could be used for upcoming versions. More generally, it allows back and forth conversions of UD and more classical subcategorization-based dependency annotation schemes.

Acknowledgments

We thank our three reviewers for valuable remarks and corrections. This work is supported by the French National Research Agency (ANR) with the project NaijaSynCor.

[10] The conversion of chains of auxiliaries (*would have been done*) to a surface syntax-based representation (would `–antiaux>` have `–antiaux>` been `–antiaux>` done) is presently problematic in UD 2 because all auxiliaries depend on the lexical verb. This suggests enriching the UD annotation either in the same way as proposed here in analysis A (with a `casedep` feature for a second `case` introduced by a first `case`) or by replacing the current bouquet style annotation with a chain of auxiliaries, an auxiliary depending on the auxiliary it subcategorizes.

[11] In this paper, we started from the UD annotation scheme and we have used UD's relation names. The names `case` and `anticase` could suggest that `case` has a sort of primacy on `anticase`. But `anticase` is simply the `obj` relation between a preposition and its direct complement.

References

Timothy Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL workshop on Multiword Expressions: analysis, acquisition and treatment,* Association for Computational Linguistics.

Charles Fillmore, Paul Kay, and Michael O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The case of "Let Alone". *Language*, 64:501-538.

Kim Gerdes and Sylvain Kahane. 2016. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In *Proceedings of LAW X.*

Raymond W. Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press, New-York.

Martin Jönsson. 2008. *On compositionality. Doubts about the Structural Path to Meaning*. PhD thesis, Lund University.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar, Volume 1: Theoretical Preresquistes*. Stanford University Press, Stanford.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony P. Cowie (ed.) *Phraseology. Theory, analysis, and applications*, 23-53.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC.*

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language,* 70:491-538.

V. Rosén, G. Losnegaard, K. De Smedt, E. Bejcek, A. Savary, A. Przepiórkowski, M. Sailer, and V. Mitetelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the Treebanks and Linguistic Theories conference* (*TLT*), Warsaw, Poland.

Ivan A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing*, 189-206.

Agata Savary, C. Ramisch, S. Ricardo Cordeiro, F. Sangati, V. Vincze, B. QuasemiZadeh, M. Candito, F. Cap, V. Giouli, I. Stoyanova, and A. Doucet, A. 2017. The PARSEME Shared Task on Automatic Identication of Verbal Multiword Expressions. In P*roceedings of the 13th Workshop on Multiword Expressions* (*MWE 2017*).

Maria Helena Svensson. 2008. A very complex criterion of fixedness : Non-compositionality. In Sylviane Granger and Magali Paquot (eds.). *Phraseology: An interdisciplinary perspective*, 81-93. John Benjamins, Amsterdam / Philadelphia.