# Dependency Annotation Choices : Assessing Theoretical and Practical Issues of Universal Dependencies

Kim Gerdes, Sylvain Kahane

▶ **To cite this version:**

Kim Gerdes, Sylvain Kahane. Dependency Annotation Choices : Assessing Theoretical and Practical Issues of Universal Dependencies. LAW X (2016) The 10th Linguistic Annotation Workshop : 131, 2016, Berlin, Germany. halshs-01509118

**HAL Id: halshs-01509118**
**https://shs.hal.science/halshs-01509118**

Submitted on 19 Sep 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies

**Kim Gerdes**
Sorbonne Nouvelle
ILPGA, LPP (CNRS)
kim@gerdes.fr

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

## Abstract

This article attempts to place dependency annotation options on a solid theoretical and applied footing. By verifying the validity of some basic choices of the current dependency reference framework, Universal Dependencies (UD), in a perspective of general annotation principles, we show how some choices can lead to inconsistencies and discontinuities, partly due to UD's alternation between syntax and semantics. For some constructions, we propose better suited alternative structures with a clear-cut distinction of syntax and semantics. We propose a classification of conception-oriented, annotator-oriented, and finally, treebank end-user-oriented considerations to be used in the creation of new annotation schemes.

## 1 Introduction

Every project of corpus annotation is about making choices. Astonishingly little research is actually going into this founding act of every treebank.

### 1.1 Justifications of treebank annotation

In the literature, the discussions of the considerations taken into account in treebank and annotation scheme constructions are rather scarce. Treebank guidelines commonly make do with the *'what choices'* rather than the *'why those choices'*. Justifications are given in theoretical works only, if the treebank is based on a framework. For the Prague Dependency Treebank (Böhmová et al. 2003) for example, choices are based on theoretical works of the Prague team (Sgall et al. 1986) and if adaptations have been done for the annotation proper they are stated neither on the PDT website nor in the guidelines. For the French Treebank (Abeillé et al. 2003), the annotation choices are guided by the desire to be "*compatible with various syntactic frameworks*" and "*as theory neutral as possible*" (FTB home page) notwithstanding that we do not know how this is even possible. However, this does not explain under which considerations particular choices have been done. For Universal Dependencies (de Marneffe et al. 2014, Nivre 2015), "*The goal of the typed dependency relations is a set of broadly observed "universal dependencies" that work across languages. Such dependencies seek to maximize parallelism by allowing the same grammatical relation to be annotated the same way across languages, while making enough crucial distinctions such that different things can be differentiated.*" (UD home page) This general manifest is used to justify some choices: "*Preferring content words as heads maximizes parallelism between languages because content words vary less than function words between languages*." But this is of course insufficient to justify numerous other choices that have been done (some of which we will discuss here).

If annotation guidelines of treebanks do not answer our question, studies dedicated to the analysis and comparison of treebanks do not help much more. Kakkonen (2005) is a good example of the kinds of questions investigated in such papers, which he resumes by "*What types of annotation schemes and formats are applied?*" or "*What kinds of annotation methods and tools are used for creating the treebanks?*". For instance, Ivanova et al. (2012) compare 7 dependency treebanks and identify "*a large variation across formats*". They note that "*divergent representations are in part owed to relatively superficial design decisions, as well as in part to more contentful differences in underlying linguistic assumptions*", but do not investigate further what kinds of considerations have led to such divergences. They are more interested in "*contrastive*

*studies*" and present an "*automatic conversion procedure*".

Corpus linguistics and annotation handbooks that we are aware of are also mainly presenting different annotation schemes. Kübler & Zinsmeister (2015) describe how "*the different tagsets impose different restrictions on which phenomena can be looked up in corpora*", but the same is not done for structural annotation choices and *a fortiori* no guideline for choosing the most appropriate annotation scheme is put forward.

### 1.2 Delimitations of our study

We are here interested in syntactic and semantic dependency annotations. By *dependency annotation* we mean an annotation based on a tokenization of the text in basic units (morphemes, words, multi-word expressions, …) and a labeled directed graph of relations between the tokens.

Deciding to use a dependency annotation is a choice in itself and, as every annotation choice, must be supported by different considerations which we propose to organize in three main groups:

1. **Theory-oriented considerations**: Adequacy of dependency has been proven for syntactic as well as semantic representations (Kern 1883, Tesnière 1959, Mel'čuk 1988, Hudson 2006). For instance, predicate-argument structures can be encoded by a dependency graph between lexical units, including idioms (Mel'čuk 1988, Kahane 2003, Copestake 2005, Banarescu et al. 2013).
1. **End-User-oriented considerations**: Dependency treebanks allow training of efficient parsers (Nivre et al. 2007, Bohnet 2010) and developing text generation systems (Bohnet et al. 2010) or translation system (Čmejrek et al. 2004). Specialized query systems exist but are still rather complex and difficult to use for the common linguist (Krause & Zeldes 2015). Dependency can also be used for grammar learning and language learning.[1] The usability of the resulting treebanks for the training of statistical parsers is also an important usage consideration (Schwartz et al. 2012).
2. **Annotator-oriented considerations**: Dependency structures are a light-weight annotation in terms of graph complexity (compared for example to phrase structure trees) and various ergonomic annotations tools have been developed (Gerdes 2013). Moreover, the annotators' evaluation is straightforward on dependency structures (labeled and unlabeled attachment scores, see Nilsson et al. 2007).

In this paper, we will explore the various choices that must be made when developing a dependency-based annotation, compare choices made by different frameworks (especially UD), evaluate on which considerations their annotation choices are based, and explore whether better choices could have been done with similar or other considerations.

The next sections will study some phenomena where basic annotation choices are traditionally made: Tokenization in section 2 exemplifies the choice of minimal units. Grammatical functions in section 3 exemplify labeling choices. In section 4, coordinations, prepositions, and light verbs exemplify structural choices. Section 5 presents an overview of the different considerations that can influence annotations choices. This last section can be read before the others and we will refer to it all along the paper.

## 2 Tokenization

Determining the units that constitute the base of the dependency structure, the tokens, is a central choice of the annotation scheme. In a syntactic treebank, basic units are words or lexemes, while in a semantic treebank, basic units are lexical units, including idioms which are multi-word expressions (MWEs).

### 2.1 Syntactic tokenization

Two options are possible: the tokenization can be based on theoretical considerations of *wordness* (*adequacy*)[2] (in which case each token has to be validated and possibly disambiguated before the dependency annotation can even start) or on purely formal spelling-based criteria like space and punctuation of the text (or the transcription

[1] Kahane & Osborne (2015) point out the pedagogical orientation of the Reeds & Kellogg (1877) diagrams as well as Tesnière's work whose basic goal was advances in language learning. See also Gerdes (2013), Zeldes (2016) who uses dependency annotation of a corpus for teaching syntax.

[2] These keywords refer to different considerations in annotation choices. They will be summarized in section 5.

for spoken corpora) (*simplicity*). The non-congruence between these considerations is an important problem for any kind of annotation scheme and calls for special annotation devices. The rules of what signs constitute word segmenters are language dependent. For example hyphens and apostrophes: The apostrophe is rather seen as part of the preceding word in French (*l' ami* 'the friend') and of the following word in English (*I 'm*). But as always, exceptions exist: Fr. *aujourd'hui* 'today', En. *isn't*. In any case, we recommend a purely formal tokenisation based on orthography and a few formal rules (*formalization*). A too fine-grained tokenization can be handled by a special dependency relation.[3]

## 2.2 Multi-word expressions

Suppose now that we develop a semantic treebank. If we want the token to be our basic semantic unit (*adequacy*) (choice A), we need a lexicon of MWEs, which is a very large resource (the number of MWEs is greater that the number of lexemes) the outlines of which are fuzzy and controversial. This is why we recommend a tokenization at the syntactic level with an encoding of MWEs by means of an additional annotation at the dependency level. This choice gives way to several options.

The seemingly most simple annotation is the one advocated by UD: Tokens which are part of a MWE are connected with a special dependency (called *mwe*[4]) and each token of the MWE as well as the MWE's external relations depend on one fixed (the first) token (*formalization*) (choice B) (Fig. 1).

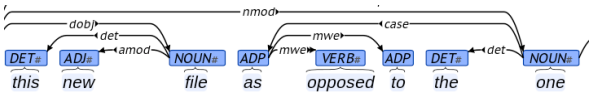


*Figure 1: 'as opposed to' as an example of a MWE, extract from UD 1.2 English*

[3] If, inversely, the spelling based units are too large, like in German N-N constructions that are written without spaces, the decomposition into semantic units requires a specific encoding mechanism. A tokenization into lexemes will need access to a lexicon, which can be costly (*concision*) and every change in the lexicon or error of tokenization implies a drastic change of the dependency structure.

[4] Additionally, UD distinguishes *compound*, *goes-with*, *name*, *foreign* for various cases of semantic units beyond the token.

Another solution is to systematically preserve the internal syntactic structure of MWEs (*level coverage*) (choice C), the majority of which have a regular syntactic pattern (Fig. 2).

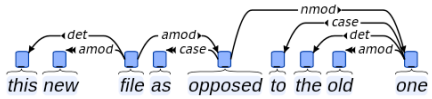


*Figure 2: compositional UD-style analysis of 'as opposed to' preserving regular syntactic labels*

Most syntacticians would agree that *as opposed to* is rather idiomatic and *as a great alternative to* isn't. The continuum between the two structures does not have a clear and consensual break-off point: *as opposed to, as relating to, as referred to, as commonly referred to, as a great alternative to,* etc.

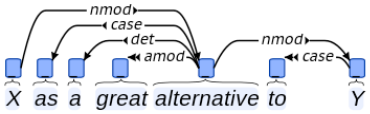


*Figure 3: compositional UD-style analysis of 'as a great alternative to'*

UD's MWE analysis therefore gives rise to a *catastrophe*, in a strictly mathematical sense of Thom's catastrophe theory (Saunders 1980), i.e. a brutal structural change in a continuum: The UD annotators have to give drastically different structures the moment they detect idiomaticity, which necessarily leads to low inter-annotator agreement, whereas the systematically compositional annotations would all look similar (*independence*) (Fig. 3).

An annotation of MWEs is compatible with the compositional structure of choice C. Two solutions are possible to add the MWE information. Choice C1: replacing the regular syntactic label with the *mwe* label (*simplicity*); or choice C2: preserving the regular syntactic label and combining it with the *mwe* label (*separability*) (Fig. 4).

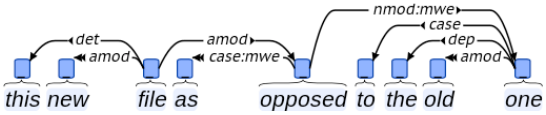


*Figure 4: complex function names of choice C2*

As most of the MWEs are either lexically or structurally non-ambiguous, obtaining C2 is not more complicated than B for the annotator (*naturalness*). Moreover, both choices C1 and C2 are structurally more informative than choice B: They can trivially and automatically be transformed into choice B, whereas the inverse auto-

matic transformation is impossible without strong resources (*transformability*). Choice C2 is the richer solution: It is possible to project C2 onto C1.

C2's additional *mwe* tag relies again on access to a MWE lexicon but the likely inter-annotator disagreement caused by the identification of MWEs exclusively consists in this additional label, no other parts of the structure are concerned (*quality*). This additional *mwe* tag can partly be added automatically, using a MWE lexicon indicating which MWE are non ambiguous. Most grammatical MWE (such as complex prepositions) can be unambiguously detected, as soon as the syntactic structure of the text is given (*minimality*).

Another advantage of the syntactically valid internal annotation of MWEs is that the transparency of the internal structure gives rise to combinatorial properties of the semantic unit for example in coordination. Consider the following example: *oneself as opposed to other selves and to everything that is "not-self."* (*fakebuddhaquotes.com*) Here the preposition *to* as part of the complex preposition is coordinated with a simple *to*, thus revealing that a more adequate analysis is to consider that *as opposed* is the MWE proper and *to* its subcategorization marker (see Fig. 4). This also causes the parser to have more similar training examples and fewer ambiguities to resolve (*precision*).

From the end user's point of view, too, the advantage of, for example, encoding the MWE *as opposed (to)* compositionally are obvious: The user of the treebank has to know only the treebank's analysis of noun and prepositional phrases to query the treebank (*readability*).

## 3 Labeling Choices

The labels of syntactic dependencies traditionally encode grammatical functions, i.e. the role the dependent plays vis-à-vis its governor and in the construction. The label can also encode categorical information (i.e. information concerning the token and not only its role). This is what UD does when they distinguish *nsubj* and *csubj*, i.e. nominal vs. clausal subjects or *nmod* and *amod,* i.e. nominal vs. adjectival modifiers. This goes against the minimality of the label set (*concision, separability*).

UD also makes the distinction between *nsubj* and *nsubjpass* (as well as *csubj* and *csubjpass*), which is a combination of syntactic and semantic information: An *nsubjpass* is a syntactic subject that does not correspond to the first actant of the verb (cf. Mel'čuk 1988, partially following Tesnière 1959[2015]: ch. 51). Maybe it would have been better to clearly separate syntax and semantics since *nsubjpass* can designate a second or third actant (*A book* ←nsubjpass– *was given to Craig* vs. *Craig* ←nsubjpass– *was given a book*). This could be done by indicating the semantic actance number, which subsumes UD's analysis and the distinction between *nsubj* and *nsubjpass*: *it* ←subj:0– *is raining* (non actancial subject), *Ann* ←subj:1– *gives Craig a book, A book* ←subj:2– *was given to Craig*, *Craig* ←subj:3– *was given a book* (*separability, transformability, level coverage*).

Redistribution between second and third actants also exists in some languages (antipassive, including dative-shift in English for some linguists (Bresnan 1981)), which cannot be encoded cleanly without introducing similar distinctions for *dobj* (Mel'čuk 1993). UD uses the label *nmod* for a dative object when it is indirect (*give a book to Craig*: *give* –nmod→ *Craig* –case→ *to*) and *iobj* when it is shifted (and direct!) (*give Craig a book*: *give*–iob→ *Craig*), which is quite counterintuitive (*intuitiveness*). Again a clear separation between syntax and semantic would be better: *Ann gives Craig a book*: *gives* –dobj:3→ *Craig*; *give* –dobj:2→ *book* vs. *Ann give a book to Craig*: *give* –nmod:3→ *Craig* –case→ *to* (*separability, adequacy, level coverage*).

## 4 Structural Choices

Orthogonally to tokens and function labels, the structure itself is matter of central choices. The basic constraint that most annotation schemes put up is the tree structure, i.e. each token has exactly one governor (including the root that can be governed by an anchor). There are many practical reasons for this choice ranging across the whole spectrum of considerations that we propose: Theoretical as well as practical, in particular as the annotation tasks get considerably harder when annotating graph structures (*simplicity, minimality*).

### 4.1 Position of the preposition

UD favors links between content words. For this reason, prepositions that mark the relation between the content words are dependents of the word they mark: *Ann talked to Craig*: *talk* –nmod→ *Craig* –case→ *to*. Consequently, every preposition is treated as a leaf of the tree, which is problematic because some prepositions are content words: *Ann talked during the play*: *talked* -nmod→ *play* -case→ *during* (*adequacy*). At first sight UD's solution seems to give the advantage of *uniformity*, but languages use compositional expressions (such as *in the (exact) middle of*, *on the (very) left of* …), which occupy the same syntactic position as prepositions while not being treated in the same way (Fig. 5). Experiments on training of different parsers (Schwartz et al. 2012) also show that prepositions as heads give higher accuracy than when they are nominal dependents (*learnability*).

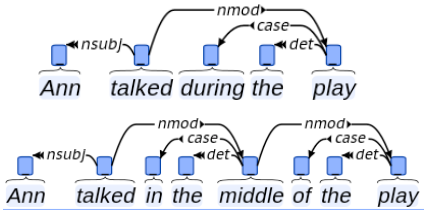


*Figure 5: analyses of both a simple and a complex semantically full preposition in UD 1.2 English*

Because of the high degree of compositionality and modifiability of expressions like *in the middle of*, UD chooses to encode these complex "prepositions" compositionally and not as MWEs (see 2.2) and consequently not as prepositions. Indeed, *middle* is treated as a content word, depends on the verb, and governs a complement (Fig. 5). In other words, the catastrophe that UD avoids in treating all prepositions uniformly is just relegated to the border between simple prepositions (such as *during*) and compositional prepositional expressions (such as *in the middle of*).

Even *universality* cannot be ensured because parallel expressions can be expressed differently in other languages. For example, the English structure of *in the middle of* will not be easily comparable to its German adverbial counterpart *mitten* and both constructions receive quite different structures (Fig. 6).

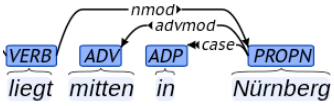


*Figure 6: German adverbial construction translating the English "in the middle of", extract from UD 1.2*

To avoid a catastrophe, it is better to preserve the syntactic structure and to have the preposition as the head of its complement (we might call this function *pobj*, for *object of the proposition*) (Fig. 7). This solution is equivalent to UD's solution (each one can automatically be transformed into the other), but our solution avoids a catastrophe (*uniformity*).

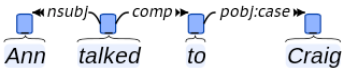


*Figure 7: Proposed analysis of a phrasal verb*

An additional label on the dependency (and/or on the preposition node) can indicate that the preposition is empty and only serves as a subcategorization marker (Fig. 8). This solution is now richer than UD's solution since it distinguishes content and phrasal verb prepositions (*transformability, level coverage*, *separability*).

Even for the comparison of languages and paraphrases it will be better (*universality*): For instance, *X cause Y* and *Y because X* will be much more parallel with our analysis, the synonymous content words *cause* and *because* being linked to their two actants in both constructions.[5]

### 4.2 Coordination

Structural analyses that go beyond the tree structure are frequently encountered for constructions involving coordination. Paradigmatic relations between words are orthogonal to government-dependent links (Tesnière 1959, Blanche-Benveniste 1990, Gerdes & Kahane 2009) and are difficult to encode in simple tree structures. Moreover, paradigmatic relations are involved in complex deletion rules that some syntactic frameworks analyze with empty nodes, something that dependency theory traditionally attempts to avoid.

[5] As already stated by Mel'čuk (1988), paraphrasing is a particular case of translation (i.e. intra-language translation) and an analysis cannot be universal (and translation-invariant) without being paraphrasing-invariant.

A simple coordination such as *we have apples and bananas* already gives rise to various links that could be encoded in the annotation scheme: *have → apples, have → bananas, apples → and, apples → bananas, and → bananas.* The direction of some links is also open for debate, in particular *apples → bananas* and *and → bananas*. From a theoretical standpoint we would like to obtain the complete graph (Gerdes & Kahane 2015), but practical considerations of annotation and query opens the question whether the structure can be simplified to a tree without loosing important information (*minimality, readability*).

Mel'čukian surface syntax handles the coordinative conjunction as a head of the second conjunct which gives an asymmetrical analysis *apples → and → bananas* (Mel'čuk 1988). UD proposes both a complete graph and a reduced tree structure. For the reduced tree structure, UD selects the paradigmatic relation *apples → bananas* consistent with UD's basic concept of relegating function words to lower positions in the tree (although a word like *and* is far from being semantically empty) (*adequacy*). This choice also allows for a consistent analysis of the frequent cases where the coordinating conjunction is absent (*uniformity*).

It remains to choose where to attach the coordinating conjunction, on the head of the first or of the second conjunct? Here UD selects the first conjunct, without further justification. Where to attach the conjunction may not be relevant from a semantic point of view, but syntactically, *and bananas* clearly is a phrase (that can be separated prosodically and also added by a second speaker in a dialogue) whereas *apples and* does not have these properties. Here the *adequacy* and the *level coverage* considerations should make us prefer the opposite choice of UD.

Shared dependents, as in *we have rotten apples and bananas,* cannot be cleanly expressed with a simple tree structure (some frameworks attach the shared adjective on a different level, e.g. the coordinating conjunction; others like Mel'čuk have specialized function labels to indicate the scope) and UD offers to either not encode the scope of the adjective (*precision*) or to upgrade to a graph structure (*adequacy*) (Gerdes & Kahane 2015).

Contrarily to the Dutch CGN corpus that skips reparanda (Schuurman et al. 2003), UD proposes to encode them with a special *reparandum* link that goes from the "correct" *repair* part to the "incorrect" *reparandum* (*text coverage*), but in the opposite direction of the *conj* link that goes from left to right. This is again a semantic choice where the semantically peripheral elements are relegated to the lower parts of the tree. Gerdes & Kahane (2009) (following Blanche-Benveniste 1990), however, show that there is a continuum between elaboration and disfluency with frequent borderline cases like "*I saw a room, a bright room, a room with red lights...*", which makes them postulate the same dependency analysis for all those cases ranging from coordination to disfluency. Thus, in UD, we again have an annotational catastrophe: The direction of the central paradigmatic link between the conjuncts depends on whether the annotator considers the second conjunct to be a correction of the first (*independence*).

The UD guidelines also include the analysis of non-constituent coordination (NCC) as in *Marie went to Paris and Miriam to Prague* by means of a specific *remnant* link that connects the elements that play the same role in both conjuncts: *Marie* –remnant→ *Miriam* and *to Paris* –remnant→ *to Prague*. Again, to prioritize on these links in a manually corrected annotation setting is a reasonable choice from a *minimality* and *naturalness* point of view.

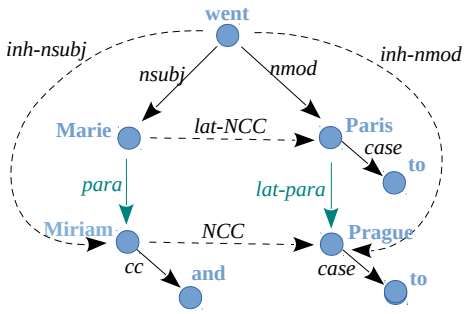


*Figure 8: NCC structure in "*Marie went to Paris and Miriam to Prague*" following Gerdes & Kahane (2015), prepositions analyzed in UD style*

However, *Miriam to Prague* also forms a constituent according to autonomy criteria (prosody and stand-alone properties, see Gerdes & Kahane 2013). This constituent is disconnected in the UD analysis and it would be preferable, if we allowed ourselves a graph structure, to add a link between *Miriam* and *to Prague.* We claim that

this link is more visible on the surface than UD's *remnant* links.

Gerdes & Kahane (2015) show a complete schema of relations[6] that arise in an NCC from which one has to choose a possible tree structure, with *para* and *lat-para* being UD's *remnant*, *NCC* and *lat-NCC* linking the constituents involved in the same unique coordination (Fig. 8). The idea is that, from a theoretical point of view, *para* and *NCC* are the primitive links, while *lat-para* and *lat-NCC* are "lateral" links, inherited from them and "symmetrizing" the structure. Nevertheless, it is a symmetric problem to automatically compute the NCC links (*NCC* and *lat-NCC*) from the paradigmatic links (*para* and *lat-para*) or the inverse: computing the paradigmatic links from the NCC links. However, only UD's choice of *remnant* links results in a tree structure and is thus preferable (*transformability, naturalness*).

Concerning coordination, we can sum up our observations by noting that in general the UD choices are well-founded in the proposed considerations with few exceptions, but these considerations are not made explicit.

### 4.3 Light verb constructions

A governed preposition (like *to* in *talk to Craig*) can be seen as reification of the semantic link between the verb and its actant. This tendency to reify semantic relation is not limited to government: copula or light verbs have the same role:

*a red book* vs. *the book is red*

*Ann's slap on Craig*
vs. *Ann gave Craig a slap*
vs. C*raig got a slap from Ann*

UD favors the semantic relations in all the cases of prepositions and copula, but not for light verbs. As explained in Nivre & Vincze (2015), the predicative noun is encoded as the *dobj* of the light verb in all UD treebanks, which is incoherent with the analysis of the copula (as a dependent of the predicative noun or adjective) (*uniformity*) and actants of the light verb construction are linked to the verb, which is incoherent with UD principles because the predicative noun is the content word (*adequacy*). Fig. 9b gives an analysis coherent with UD principles, to be compared with the present analysis of UD.

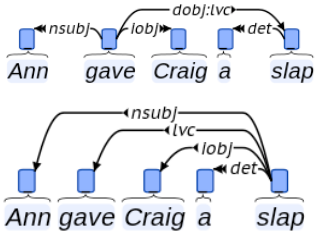


*Figure 9: a. UD analysis of a light verb construction*
*b. coherent analysis of a light verb construction*

Of course, such an analysis is also problematic because the frontier between light verbs and content verbs is quite fuzzy (see for instance the very rich classification of support verbs in Mel'čuk (1998), cf. *feel fear* vs. *shake with fear*).

We then recommend maintaining the present annotation of LVC, which is similar to the syntax-based annotation of prepositions we have recommended. But to avoid a catastrophe, the same analysis should be used for the predicative construction: *book* ← nsubj- *is* -dobj:lvc→ *red*.

## 5 Overview of considerations about annotation choices

In this section, we propose to categorize the different types of considerations that we exemplified in the preceding sections. There are three stages in the development of a resource: conception, realization, and usage.

### 5.1 Conception-oriented consideration

The first decision concerns the kind of linguistic information we want to develop in our annotation. According to our theoretical goals, our annotation must respond to the following considerations:

A1. **Adequacy**: Our annotation must be as adequate as possible given our theoretical framework and the criteria validating a correct analysis.

A2. **Uniformity**: Similar constructions must be annotated in similar ways. Catastrophes must be avoided.

A3. **Level coverage**: Our annotation must be as informative as possible and must cover the maximum of linguistic levels. It can be costly to develop a too fine-grained annotation, but for a comparable cost, the more precise annotation must be chosen.

[6] The graph also includes the "inherited" links *Miriam* ←inh-nsubj– *went* and *went* –inh-nmod→ *to Prague*, which also undergo semantic and restrictional selection (see Tesnière 1959[2015]: ch. 143).

A4. **Text coverage**: Our annotation must cover the maximal range of relevant data. In terms of dependency annotation, it means that the graph must be as connected as possible. A text is cohesive and for instance many relations may not be limited to sentence boundaries.

### 5.2 Annotator-oriented considerations

The realization of a treebank supposes an annotation stage, but also some steps of validation of the annotation, as well as an easy maintenance and expansion of the treebank.

According to the need for efficiency in the annotation process, our annotation must respond to the following considerations:

B1. **Formalization**: Annotation criteria must be well formalized in order to avoid inter-annotator disagreement and to speed up annotators' decisions. A good formalization also means that part of the annotation process can be computer-assisted, by an automatic pre-annotation or by a tool pointing out inadequate annotations.

B2. **Simplicity**: The annotation process must be as simple as possible and complex or open decisions must be avoided. In particular, a tree structure can be preferred because each token has exactly one governor (except the root), which also enables an economic encoding (tabular, CoNLL) and a faster search.

B3. **Minimality**: The annotation can be enriched automatically (by deterministic and local rules) if it contains all information and all distinctions we want to make. It means that the annotation delivered by the annotators must be as minimal as possible to avoid useless work. Again, a tree structure can be preferred because, for a connected graph, a tree has the minimal number of links.

B4. **Concision**: Not only the annotation itself must be minimal but information needed to annotate must also be minimized. Tag sets as well as the guideline must be concise. Consultation of an external resource (for instance, a lexicon of multi-word expressions) must be avoided unless it is automatic and at no cost for the annotator.

B5. **Naturalness**: Annotators are humans and some decisions are easier than others for humans. Paradoxically, some high-level decisions, close to semantics for instance, can be easier than some low-level decisions, that would be much easier for a machine.

B6. **Separability**: The annotation can involve information of different levels. The choice between a unistratal annotation (combining different levels) and a multistratal annotation (separating everything that can be separated) must be made. As long as the size of the tag set remains reasonable, it could be more efficient to combine, but choosing between $n$ tags and then $m$ tags is quickly faster than choosing between $n \times m$ tags.

B7. **Independence**: A change of annotation in a particular level must not drastically affect other levels of annotation.

B8. **Intuitiveness**: Annotation is labeling. Label terms must be intuitive. Terminology must be coherent with traditional uses.

### 5.3 End User-oriented considerations

An annotation project must be aware of the applications of the developed resource. Different goals can be considered:

- **Theory**: Annotating a corpus following a particular framework can be a means of proving the adequacy of the theory and evaluating its coverage.
- **NLP**: Many tools can be developed from a treebank, in particular using machine learning methods.
- **Pedagogy**: The annotation itself can be a good exercise to practice linguistics. And an annotated corpus can be a source of knowledge for learners (and other researchers).

According to our practical goal, our annotation must respond to the following considerations:

C1. **Quality**: The annotation must be reliable. In particular inter-annotator agreement must be as high as possible.

C2. **Precision**: The annotation must be fine-grained enough for the expected applications. But too much precision is unnecessary and removing a distinction (e.g. the categorical distinction between French *des* DET vs. *des* PREP+DEP) can speed up the annotation process and lower the error rate.

C3. **Learnability**: An annotation scheme is preferable if it gives higher accuracy when used for training a statistical parser. This point is strongly dependent on the state of the

art of statistical parsing as well as on the size of the developed resource.

C4. **Readability**: The annotation must be easily interpretable by a user by a direct reading or via a query system.

C5. **Universality**: The annotation must not be too specific to a particular language or genre in order to allow extrapolation to other corpora (especially under-resourced languages) and comparisons. This concerns also spoken corpora and sign languages.

C6. **Transformability**: Annotation standards must be developed. But it is unproblematic to develop a new annotation if it can be transformed into other standards. It is essential to preserve inter-operability of resources and tools.

This list does not close the considerations taken into account. We have focused on scientific considerations, but in the end choices are political. For questions of visibility, availability of tools and guidelines, and perspective of richer collaborations, many teams choose to use the most visible annotation styles, which is a reasonable choice.

## 6 Conclusion

Every project of treebank development needs to make choices between different possible annotations. Conceptualizers of the treebank generally expose the general principles that underlie the main choices. These principles reduce the space of possible choices but as soon as we get into the details, several options remains possible, many particular choices are not argued for and it is not easy to know what considerations have at last been decisive.

In this article we concentrated on the UD annotation choices, refuting some and corroborating others based on our list of principles. This list as well as the corresponding example discussions might prove useful for future treebank development choices, in this or an extended format.

## References


Abeillé, Anne, Lionel Clément, and François Toussenel. "Building a treebank for French." *Treebanks*. Springer Netherlands, 2003. 165-187.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. "Abstract meaning representation for sembanking." *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. 2013.

Blanche-Benveniste, Claire, et al. *Le français parlé (études grammaticales)*. Sciences du langage, 1990.

Böhmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. "The Prague dependency treebank." *Treebanks*. Springer Netherlands, 2003. 103-127.

Bohnet, Bernd. "Very high accuracy and fast dependency parsing is not a contradiction." *Proceedings of the 23rd International Conference on Computational Linguistics*. ACL, 2010.

Bohnet, Bernd, Leo Wanner, Simon Mille, and Alicia Burga, A. "Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer." *Proceedings of the 23rd International Conference on Computational Linguistics*. ACL, 2010.

Bresnan, Joan. "An approach to Universal Grammar and the mental representation of language." *Cognition, 10(1)*. 1981. 39-52.

Copestake, Ann. "Slacker semantics: why superficiality, dependency and avoidance of commitment can be the right way to go." *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, 2005.

Čmejrek, Martin, Jan Hajič, and Vladislav Kuboň. "Prague Czech-English dependency treebank: Syntactically annotated resources for machine translation." *Proceedings of the 10th Conference of the European Association for Machine Translation*. 2004.

Gerdes, Kim and Sylvain Kahane. "Defining dependencies (and constituents)." *Frontiers in Artificial Intelligence and Applications, Volume 258: Computational Dependency Theory*, 2013. 1-25.

Gerdes, Kim. "Collaborative Dependency Annotation." *Proceedings of the 2nd Confenrence on Dependency Linguistics,* 2013.

Gerdes, Kim and Sylvain Kahane. "Speaking In Piles: Paradigmatic Annotation Of French Spoken Corpus." *Proceedings of the 5th Corpus Linguistics Conference*, Liverpool. 2009.

Gerdes, Kim and Sylvain Kahane. "Non-constituent coordination and other coordinative constructions as dependency graphs." *Proceedings of the 3rd international conference on Dependency Linguistics*. 2015.

Hudson, Richard. *Language Networks: The New Word Grammar*. OUP Oxford, 2006.

Ivanova A., Oepen S., Øvrelid L., Flickinger D. "Who Did What to Whom? A Contrastive Study of Syntacto-Semantic Dependencies." *Proceedings of the 6th Linguistic Annotation Workshop (LAW VI),* ACL, 2012.

Kahane, Sylvain. "The meaning-text theory." *Dependency and Valency.* An International Handbook of Contemporary Research 1. 2003. 546-570.

Kahane, Sylvain, and Timothy Osborne. "Translators' introduction." In Tesnière Lucien, *Elements of structural syntax*. 2015. xxix–lxxiv.

Kakkonen, Tuomo. "Dependency Treebanks: Methods." Annotation Schemes and Tools. *Proceedings of the 15th Nordic Conference of Computational Linguistics*. 2005.

Kern, Franz. *Zur Methodik des deutschen Unterrichts*, Nicolai, Berlin. 1883.

Krause, Thomas and Zeldes, Amir "ANNIS3: A New Architecture for Generic Corpus Query and Visualization." *Digital Scholarship in the Humanities, 31(1)*. 2015. 118-139.

Kübler, Sandra, & Zinsmeister, Heike. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Publishing.

De Marneffe, Marie-Catherine, et al. "Universal Stanford dependencies: A cross-linguistic typology." *Proceedings of 9th International Workshop on Language Resources and Evaluation*. 2014.

Mel'čuk, Igor. *Dependency syntax: Theory and Practice.* SUNY Press. 1988.

Mel'čuk, Igor. "The inflectional category of voice: towards a more rigorous definition." *Causatives and transitivity, 23*. 1993.

Mel'čuk, Igor. "Collocations and lexical functions." In A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications.* Clarendon Press, Oxford. 1998. 23-53.

Nilsson, Jens, Sebastian Riedel, and Deniz Yuret. "The CoNLL 2007 shared task on dependency parsing." *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, 2007.

Nivre, Joakim, et al. "MaltParser: A language-independent system for data-driven dependency parsing." *Natural Language Engineering, 13(2)*. 2007. 95-135.

Nivre, Joakim. "Towards a universal grammar for natural language processing." *Computational Linguistics and Intelligent Text Processing*. Springer International Publishing. 2015. 3-16.

Nivre, Joakim, and Veronika Vincze. 2015. "Light Verb Constructions in Universal Dependencies." Technical report, Parseme workshop.

Reed, A. and Kellogg B. *Higher Lessons in English: A Work on English Grammar and Composition*. Clark and Maynard, New-York. 1877.

Saunders, Peter T. *An introduction to catastrophe theory*. Cambridge University Press, 1980.

Schwartz, Roy, Omri Abend, and Ari Rappoport. "Learnability-Based Syntactic Annotation Design." *Proceedings of Coling*. 2012.

Schuurman, Ineke et al. "CGN, an annotated corpus of spoken Dutch." *Proceedings of 4th International Workshop on Language Resources and Evaluation*. 2003.

Sgall, Petr, Eva Hajicová, and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Kluwer, Dordrecht. 1986.

Tesnière L. 1959. *Éléments de syntaxe structurale.* Paris: Klincksieck [transl. by Osborne T., Kahane S. *Elements of structural syntax.* Benjamins, 2015].

Zeldes, Amir. "The GUM corpus: creating multilayer resources in the classroom." *Language Resources and Evaluation.* 2016.1-32.