Mischievous Nominal Constructions in Universal Dependencies

Nathan Schneider **Amir Zeldes**
Georgetown University
{nathan.schneider, amir.zeldes}@georgetown.edu

Abstract

While the highly multilingual Universal Dependencies (UD) project provides extensive guidelines for clausal structure as well as structure within *canonical* nominal phrases, a standard treatment is lacking for many "mischievous" nominal phenomena that break the mold. As a result, numerous inconsistencies within and across corpora can be found, even in languages with extensive UD treebanking work, such as English. This paper surveys the kinds of mischievous nominal expressions attested in English UD corpora and proposes solutions primarily with English in mind, but which may offer paths to solutions for a variety of UD languages.

1 Introduction

Universal Dependencies (UD; Nivre et al., 2016, 2020; de Marneffe et al., 2021) is a framework describing morphology and dependency syntax cross-linguistically. It establishes common labels and structural constraints for annotating data, comparing languages, and training and evaluating parsers.

This paper, intended for readers familiar with UD (specifically, Basic Dependencies in version 2), addresses what we see as a significant shortcoming of the current guidelines: "mischievous" nominal structure—roughly, constructions that form noun phrases beyond the canonical components of determiner or possessive, adjective modifier, noun compound modifier, head noun or pronoun, modifier PP, and modifier clause. Many of these are productive but narrow constructions forming multiword names, dates, measurements, and compound-like structures.

Such expressions often buck ordinary restrictions on NP structure: Kahane et al. (2017), for instance, note that "most languages have particular constructions for named entities such as dates or titles.... These subsystems are in some sense 'regular irregularities', that is, productive unusual constructions." In other words, names and dates often do not fit the mold of other noun phrases, though as we will show below, the issues they raise pop up in other environments too. For many of these mischievous constructions, the existing UD syntactic relations are inadequate, or inadequately described, and corpora are widely inconsistent as a result—in some cases within a single treebank or between treebanks in the same language.

Many of the issues presented below have been discussed at length within the UD community but without any definitive resolution. Our goal is to consolidate the discussion and argue for a coherent approach (or set of alternatives) based on careful analyses of English constructions across a range of text types.[1] To minimize added complexity to the UD scheme, our proposals are conservative, focused on clarifying boundaries between existing labels and in some cases proposing new subtypes (which, though language-specific, may be adapted to other languages). While we refrain from proposing new universal relations that would force extensive editing across languages to maintain validity, we welcome feedback on related phenomena in other languages. Although our analysis is focused on English, we believe that similar reasoning applies to a range of other languages which cannot be adequately examined here due

[1] Some short examples in this paper come from introspection, while longer examples and statistics are taken from the English Web Treebank (UD_English-EWT; Silveira et al., 2014), and UD_English-GUM (Zeldes, 2017) or UD_English-GUMReddit (Behzad and Zeldes, 2020), which together cover a broad spectrum of spoken and written genres and writing styles.

to space reasons; we hope that guideline discussions in those languages will benefit from the analyses below.[2]

2 Name Descriptors

We turn first to proper names, especially names of persons, and the constructions by which a speaker can elaborate on a nominal referring expression.

(1) a. I met Gaspard Ulliel.
b. I met Gaspard Ulliel, the French actor.
c. I met the French actor, Gaspard Ulliel.

(2) a. I met French actor Mr. Gaspard Ulliel.
b. *I met French actor.
c. *I met the Mr. Gaspard Ulliel.

How are these handled in UD? The **`flat`** relation comes into play for open-class expressions with no clear syntactic head, canonically including personal names like *Gaspard Ulliel*. A flat structure, by convention, is represented in UD by designating the first word as the head of each of the subsequent words, which attach to it as **`flat`** (a "bouquet" or "fountain" analysis).

The trouble is that referring expressions may contain descriptors beyond personals. Following the Cambridge Grammar of the English Language (CGEL; Huddleston and Pullum, 2002), we distinguish two types of pre-name descriptors in English: An **appellation** is a title that would be used to formally address somebody by social status (e.g. occupation or gender), such as ***Mr.*** *Obama* or ***President*** *Obama*. An **embellishment**[3] is a bare nominal phrase preceding the name (and appellation if there is one) describing the referent with category information like *actor*, *French actor*, or *surprise winner of the Kentucky Derby*.[4] The embellished name may have an inanimate referent, as in ***German car maker*** *BMW*.[5] In English, embellishments are characteristic of select genres such as news.[6] (2a) contains an embellishment and an appellation within the same referring expression. The current UD guidelines state:[7]

> If the two nominals participate in denoting one entity, the default relation to connect them is **`flat`** (which may also be used to connect other nodes that are not nominals). Typical examples are personal names: we can say that *John Smith* is a special type of *John* as well as a special type of *Smith*, but none of the names governs the other and either of them can be omitted. In many languages this analysis extends to titles and occupations, as in English *president Barack Obama*.

Yet the flat analysis for embellishments and appellations yields counterintuitive results. That they are bare NPs and are omissible—whereas the personal name is not, as shown by (2b)—is strong syntactic evidence that they are modifiers. Moreover, it should be intuitively obvious that *Gaspard* and *Ulliel* form a coherent unit of structure—yet under the bouquet analysis for flat structures (i.e. attaching all children to the first token), *Ulliel* would have distinct heads for *Gaspard Ulliel*, *French actor Gaspard Ulliel*, and *Mr. Gaspard Ulliel*.[8]

Further discussion in the guidelines acknowledges treating titles as **`flat`** is controversial, but explains that titles do not meet typical criteria for **`nmod`**, **`compound`**, or **`appos`**. An **`nmod`** typically receives its own independent case marking (possessive or prepositional in English). **`appos`** is limited in UD to relations

[2] An extended version of this paper (Schneider and Zeldes, 2021) contains additional recommendations regarding numbers and adverbial NPs, omitted here due to space limitations.

[3] Also called "false title", described here as a kind of apposition: `https://en.wikipedia.org/wiki/False_title`

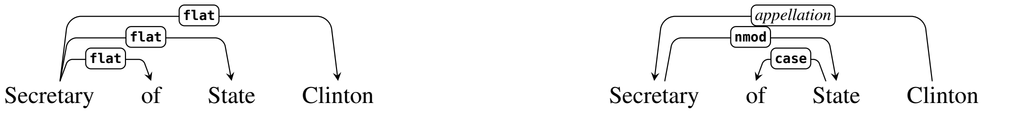
[4] An anonymous reviewer has commented on the difficulty of applying the bare nominal diagnostic in languages with different determiner systems, such as Slavic languages, Chinese, or Japanese. We fully acknowledge that equivalent constructions may look quite different in those languages, but also believe that the problems analyzed here are both substantial enough in English to merit a more detailed treatment, and common enough in other languages that the discussion is likely relevant beyond English.

[5] Thanks to an anonymous reviewer for this example.

[6] A newscaster might say, ***Surprise winner of the Kentucky Derby*** *American Pharaoh received a hero's welcome upon returning home today...*. Note the lack of an article at the beginning of the sentence.

[7] `https://universaldependencies.org/workgroups/newdoc/two_nominals.html`

[8] Note that some embellishments and appellations contain clear internal structure (e.g., ***French actor*** *Ulliel*—**`amod`**; ***Secretary of State*** *Clinton*—**`nmod`**, **`case`**). This does not pose an additional problem for the **`flat`** analysis, however: even dependents within a flat structure may host internal modifiers, as was recently clarified in the guidelines.



(a) Current flat analysis

(b) Proposed modifier structure for appellation

Figure 1: An appellation with current vs. proposed structures (several options for the relation label of the "*appellation*" dependency discussed below).

between two full NPs (or DPs, i.e. NPs including a determiner), as in (1b, 1c). And crosslinguistically, "titles do not usually behave like compounds: in German, they are not joined to the following words, as compounds are normally joined in German, and they appear at the beginning of names in both German and Hebrew, even though German compounds are head last and Hebrew compounds are head first."[9]

Nevertheless, we suggest that appellations and embellishments be removed from the flat analysis. Exactly how this could be achieved is considered below.

2.1 A Relation for Titles?

A narrow solution would be to group appellations and embellishments under the category of **titles**. As these constructions are frequent and distinctive, a subtype called **`:title`** might be appropriate, and subtyping could alleviate the concern that none of the existing top-level deprels is a perfect fit. Alternatively, a new top-level relation could be introduced. We thus begin by considering the following options:

- **`title`**, a new top-level relation
- **`compound:title`**
- **`appos:title`**
- **`nmod:title`**
- **`nmod:desc`**, a broader subtype, meant to cover additional mischievous nominals

A new top-level relation? A new top-level (universal) relation, **`title`**, presupposes that honorific titles, at least, occur widely across languages and may have idiosyncratic syntax. However, it seems possible that in some languages titles might have 'normal' syntax, and would not need such a top-level relation at all. Even for languages with conspicuous title syntax, UD relations aim to be as compact as possible; adding major labels is not done lightly, and would require waiting for UDv3, not to mention imposing costs on many treebank maintainers and requiring updates to existing tools. We therefore prefer subtyping an existing relation.

Problems with `compound:title`. In English, **`compound`** dependent nouns too are bare (lack a determiner of their own), similar to appellations and embellishments, suggesting a subtype **`compound:title`**. In fact, there is prior art in UD: Finnish UD documents the label **`compound:nn`** for appellations.[10]

However, there are important differences that suggest compound nominals (at least in English) and titles are two different beasts. While the definition of **`compound`** is quite vague, its applicability to modifiers of nouns is clearest in determinative compounds, either where both the head and modifier are part of a multiword proper name like *Washington Post*; or where the head denotes a kind (usually, a noun that could be made either definite or indefinite) which is restricted by the modifier, e.g. *cake flavors*. Often such non-name combinations could be paraphrased with a possessive or prepositional construction if used literally (*flavors of cake*); and often compounds behave like complex words and may become lexicalized as idiomatic multiword expressions. By contrast, appellations and embellishments of proper name heads nonrestrictively add information about an entity and might be paraphrased with "who is" or an appositive (*French actor Gaspard Ulliel* → *Gaspard Ulliel, the French actor / Gaspard Ulliel, who is a French actor*).

Morphosyntactic evidence also weighs against the compound analysis: English compound modifiers are very rarely plural, even when denoting multiple items—whereas appellations, embellishments, and appositives agree in number with their referent:

(3) a. **Presidents** Obama and Biden [appellation]

[9] `https://universaldependencies.org/u/dep/flat.html#some-further-notes-on-relations-for-names`

[10] `https://universaldependencies.org/docs/fi/overview/specific-syntax.html#appositions-and-appellation-modifiers`

b. French **actors** Ulliel and Marceau[11] [embellishment]
c. Sam and Isaac, my **brothers** [`appos`]
d. ***eggs** carton(s)

Cartons of/for multiple eggs are *egg cartons*, stripping the plural ending from the compound modifier.[12] If appellations and embellishments were special cases of the English compound construction we would expect them to resist pluralization as well, but this is not the case (3a, 3b).

Problems with `appos:title`. Part of the practical motivation for the `appos` relation is to express a semantic notion of equivalence between referring expressions, such that an information extraction system could strip out supplementary information when matching names against entities in a knowledge base. Thus *French actor Gaspard Ulliel, my hero since childhood, won an Oscar* could be simplified to *Gaspard Ulliel won an Oscar* by removing `appos` and `appos:title` dependents. From an argument structure perspective, `appos` is characterized by not adding participants to valency frames, i.e. *Gaspard Ulliel* and *my hero since childhood* both instantiate the subject of *won*.

On the other hand, `appos` is already rather complicated (spelled out in detail below, §2.3). While embellishments are sometimes categorized as appositions, there is a lack of universal agreement that appellations and embellishments qualify as appositive modifiers; other sources (e.g., Ruppenhofer et al., 2016, p. 77) view the name rather than the embellishment as the appositive phrase.

Intermediate Proposal: a subtype of `nmod`. The rationale here is that `nmod` is the most general relation for nominals modifying other nominals. (It already has subtypes, including `nmod:poss` for possessive modifiers and `nmod:tmod` for temporal modifiers.) In English, plain `nmod` dependents have case marking or prepositions, but the subtyping can signal a morphosyntactically exceptional construction, as is already the case with prepositionless `nmod:tmod`.

If we target only titles, then `nmod:title` is the least objectionable solution narrowly tailored for embellishments and appellations, given that (a) `nmod` already has other subtypes, (b) this would avoid confusion with dominant uses of `compound` and `appos`, and (c) implementing a new universal relation across treebanks would be onerous, but treebanks are allowed flexibility to diverge and innovate with subtypes. On the other hand, there are a number of other 'mischievous' adnominal constructions requiring a solution, which suggests that a subtype focusing only on titles may be too narrow, motivating a more general name fitting other types of descriptive modifiers, for which we will propose a new relation (called `nmod:desc`).

2.2 Other Special Types of Nominal Modification

The above discussion is limited to appellations and embellishments that precede a name. But other, less frequent constructions bear some resemblance to these:

(4) Post-name bare nominal modifiers:
a. 11-year-old Draco, **scion of the Malfoy family**, was sorted into Slytherin.
b. Oedipus, **King of Thebes**

(5) First or second person pronoun plus noun:[13]
a. We **pilots** deserve a pay raise.
b. You **guys** deserve a pay raise.[14]

In (4, 5), the bolded nominal phrase can be omitted while its head (underlined) cannot. (4a) can be considered a post-head embellishment, and (4b) a post-head appellation. The construction seen in (5),

[11] It is unclear whether non-coordinated names referring to multiple individuals could license plural embellishments via semantic number agreement: *An argument broke out between married actors Brad and Angelina / ?married actors Brangelina / ?British comedians Monty Python*.

[12] For exceptional pluralized modifiers in Germanic compounds see also Fuhrhop (1996).

[13] Elsewhere the pronouns are analyzed as determinatives (Huddleston and Pullum, 2002, p. 374), but we deem it impractical to extend `det` to include such specialized uses of personal pronouns.

[14] The expression *you guys* has been conventionalized in some dialects as a gender-neutral second person plural.

headed by a pronoun, is a cousin of the pre-head embellishment, as shown by the third person paraphrase of (5a): ***pilots*** *Earhart and Lindbergh*. A broad relation `nmod:desc` for the special cases seen above as well as appellations and embellishments would separate them from the `appos`, `compound`, and `flat` cases while covering sufficient ground to merit its inclusion.

2.3 More on Appositives

A classic example of an appositive appears in (6). The appositive phrase, *my brother*, is a nonrestrictive full NP descriptor of *Sam*. It is syntactically omissible, and could in fact replace its head as they share the same referent. A similar phenomenon appears in (7), where an indefinite NP ascribes a property to *Sam*:

(6) Sam, **my brother**, is very tall.

(7) Sam, **a musician**, is very tall.

The current definition of the `appos` relation establishes the following criteria:

(8) An appositive (`appos`) must be
a. a full NP
b. modifying an NP in a reversible fashion (modulo punctuation)
c. to the right
d. with no intervening words.[15]

While appositive phrases are often separated by commas or parentheses, this is not a strict requirement, and of course spoken language has no commas. We understand the definition to also include:

(9) a. my brother **Sam**
b. the color **purple**
c. the word "**terrorist**"
d. the play ***Much Ado About Nothing***

Cases resembling appositives in some but not all of the above respects require clarification. The bare modifiers discussed above are sometimes considered appositives, but UD excludes them with criterion (8a). (10) satisfies criteria (8a, 8c) but not (8b, 8d), whereas (11) satisfies (8a, 8b, 8d) but fails (8c):

(10) "Maybe she really does just need a little space. . . ," Amy said, **ever the optimist**.[16]

(11) **A new Pakistani leader**, he is intent on instituting reforms.

There seem to be two ways forward:

- Relax `appos` criteria either in general or in a subtype. In particular, relaxing (8b–8d) would allow `appos` to cover (10, 11). This would contrast with `nmod:desc` suggested above, which covers bare nominal modifiers.
- Maintain the `appos` criteria in (8), and classify examples such as (10, 11) as `dislocated`. These constructions are not quite classic dislocation constructions,[17] but they could be treated as if removed from their normal apposition location.

In the interest of maintaining the status quo for appositions, we favor the latter solution and recommend using `dislocated`.

		head	modifier optional?	invertible?	agreement?	type	relation
(2a)	actor Ulliel	R	Ulliel	*Ulliel, actor	actors Ulliel and Marceau	name (head)	←**nmod:desc**
§2	President Obama	R	Obama	*Obama, President	Presidents Obama and Biden	name	←**nmod:desc**
§3.1	Church Street	R	*Street / the street	*Street, Church	Church and River Streets	name	←**compound**
(12)	Lake Michigan	L	*Lake / the lake	*Michigan, Lake	Lakes Michigan and Ontario	name	**compound**→
(14)	Figure 4	L	*Figure / the figure	*4, Figure	Figures 4 and 5	name w/ num	**nummod:name**→? **compound**→? **nmod:desc**→?
(13)	Firefox 58.0	L	Firefox	*58.0, Firefox	*Firefoxes 58.0 and 59.0	name w/ num	**nummod:name**→? **flat**? **nmod:desc**→?
§3.7	London, UK	L	London	*UK, London	*Londons, UK and Ontario	name	**nmod:npmod**→
	Joe Biden	–	(flat)	(flat)	*Joe and Jill Bidens	name	**flat**
(6)	my brother Sam	L	my brother	Sam, my brother	my brothers Sam and John	name (mod)	**appos**→

Table 1: Constructions involving names and their syntactic properties.

3 Further Issues with Names

3.1 Syntactically analyzable proper names

Several other aspects of the syntax of names need to be addressed. The syntactic properties of many of the constructions at issue are summarized in table 1. We begin by underscoring UD's policy of analyzing the internal structure of names with ordinary syntax where possible, regardless of the semantic status of the name. For example, *Church Street* is analyzed with **compound**; and *New York City* consists of an adjective which modifies a noun (**amod**), which in turn modifies another noun (**compound**).[18]

3.2 Cardinal directions

Cardinal direction modifiers of nouns (*north*, *northeast*, etc.) are annotated inconsistently in English UD corpora. Based on the tagging tradition of LDC corpora, these should be treated as nouns unless they bear overt adjectival morphology (*northern*, etc.). Cardinal direction nouns premodifying nouns should therefore attach as **compound**, whether the expression is a proper name (*North Carolina*) or not (*north coast*). When multiple parts of a cardinal direction term are separated by a space or hyphen, they are joined with **compound**: e.g. *north east* 'northeast'.

3.3 Names beginning with an entity type

Many proper names incorporate a transparent entity type. In *the Thames River*, the name is constructed as an ordinary endocentric compound, with the entity type last and serving as the head and an identifier as the modifier.[19] But *the River Thames* (along with the other examples in (12)) poses a problem as the order is reversed:

(12) a. Mount Fuji
b. Fort Knox
c. Lake Michigan
d. the River Thames

It can be argued that the head in (12d) is then *Thames*, as *River* can be omitted: *the Thames* (Huddleston and Pullum, 2002, pp. 519–20). However, this omission of the entity type could be viewed as a shortening

[15] An exception to this constraint is already found in languages with so-called Wackernagel particles, such as Classical Greek or Coptic, which appear in the second position in the sentence and can interrupt any phrase or dependency; see Zeldes and Abrams (2018).

[16] *The Body in the Casket: A Faith Fairchild Mystery*, Katherine Hall Page, 2017

[17] The preferatory appositive in (11)—which features a description followed by a definite NP, and would be perfectly at home in a newspaper—is not to be confused with hanging topic left-dislocation with a pronoun referring back to the dislocated element, as might be uttered in conversation: *My dad, he is always running late.*

[18] Previously, POS tags in the English treebanks followed Penn Treebank tags and treated all content words within a proper name as PROPN, but this was changed in v2.8; PROPN is now limited to nouns.

[19] Other place names headed by an entity type and exhibiting ordinary syntax include *Mirror Lake*, *Ford's Theatre*, and *the Dome of the Rock*.

not unlike reducing *Fenway Park* to *Fenway* on the assumption that the speaker is able to identify the referent based on the more specific part of the name. Such shortenings will vary in felicitousness depending on the particular name and context. (Plain *Michigan* does not refer to the same thing as *Lake Michigan*.)

Note also that the name-initial entity types may be pluralized when grouping together multiple entities of the same type, which distinguishes them from flat structures or typical compound modifiers and suggests they may be heads: ***Lakes** Michigan and Ontario* (cf. *Mirror and Swan **Lakes***). This fits with the expected semantics, as noun-noun compounds tend to be headed by the superordinate category, and historically it is possible that the construction is in fact a remnant of left-headed compounding from Romance place names, possibly from Norman toponym patterns (English *Mount* X, French *Mont*-X, e.g. *Mont-Saint-Michel*).

We therefore consider the examples in (12) as inverted (left-headed) compounds.[20] The identifier can attach to the entity type as **`compound`** to reflect the inverted word order in these kinds of names.

3.4 Numbered entities

Numbers can also figure into names. They can disambiguate multiple of a series of related entities named by a proper noun, as in (13). These are appendages to a proper name, syntactically omissible (with a resulting broadening of meaning), and could be treated as modifiers. Numbers can also follow an entity type, as in (14).

(13) a. Firefox (version) 58.0
b. Richard III
c. *Toy Story 3*
d. 1 Corinthians
e. World War II

(14) a. Figure 4
b. room 11b
c. pp. 5–10
d. subpart (e)
e. item (number) 3
f. *Symphony No. 5*

The cases in (14) use the number to identify a specific instance of the type. The entity type appears first, similar to the inverted **`compound`** examples in §3.3. It is a completely different construction from quantity modification, the predominant application of **`nummod`**, as in *3 items* (plural!) or *3%*. A morphosyntactic difference between the numeric modifier constructions in (13) and (14) is that only the latter exhibit agreement: *page 5* (one page), *pages 5–10* (multiple pages), but **Firefoxes 58.0 and 59.0*.

We see three options, each with pros and cons:

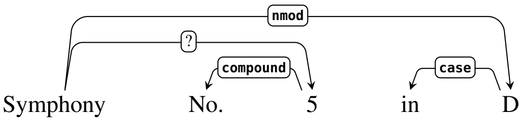
- The morphosyntactic difference notwithstanding, treat (13) and (14) as essentially the same construction, with a new relation such as **`nummod:name`** (consistent with the fact that the superordinate category **`nummod`** is currently applied to numeric modifiers generally).[21] Advantages are that (13) and (14) look very similar, and numbers are a salient property for annotators or corpus users to notice when selecting the appropriate relation. However, adding a subtype for a relatively narrow and infrequent phenomenon is questionable, and some cases are not numeric (*Level B*).
- Treat (13) and (14) as instances of more general constructions. The construction in (14) can be considered an inverted compound like *Lake Michigan* (§3.3). Flat structures could apply to the names in (13) as this construction is less morphologically transparent. This would avoid a new subtype but also may be seen as splitting hairs based on a subtle morphosyntactic criterion.
- A third option is to adopt **`nmod:desc`** for the constructions in (13) and (14). This would essentially restrict the definition of **`compound`** to substantive lexical material excluding numbering designators; **`nmod:desc`** would broadly cover miscellaneous modifiers associated with names that do not fit the more conventional constructions. This solution eclipses the similarity between *Lake Michigan* (which would remain **`compound`**) and *Figure 4*, but it perhaps avoids a counterintuitively broad application of **`compound`**. It also means that the scope of **`nmod:desc`** is a bit broader, including not just modifiers

[20] Another analysis we considered was to treat the entity type as an **`nmod:desc`** modifier, giving *Lake Michigan* the same structure as *Dr. Livingstone* or *actor Ulliel*. But the entity types in (12) seem more central to the name than titles, and are not as freely omissible, so we are not persuaded that they are modifiers.

[21] The choice of subtype parallels **`flat:name`**—an optional subtype not currently implemented in English corpora, though it is used for a number of corpora in other languages. The **`flat:name`** guidelines currently include *Formula 1* as an example; this would become **`nummod:name`** in this option.

that are secondary to the main part of a name, but also modifiers that are essential to it (just *Figure* is not a name, whereas *Ulliel* is).

(13a, 14e, 14f) illustrate a construction in which a word like *number* or *version* may precede a number to clarify that it is an identifier rather than a quantity. In modern usage this would generally remain singular even if referring to multiple items (*items number 3 and 4*), so we analyze *number* as a **compound** modifier by default, and **nmod:desc** only if plural (*items numbers 3 and 4*). "?" is provided as a stand-in for the relation between the entity type and the number given the above uncertainty:[22]



For hyphenated numeric ranges (14c), the prevailing policy in UD corpora has been to analyze the second part like a prepositional phrase *to 10*, thus an **nmod** of *5*. One of the authors takes the view that a coordination analysis would be more natural. In any event, *5* attaches to *pp.* as a modifier.

3.5 Business and personal name suffixes

Adjective-expanding suffixes like *Inc.* ("incorporated") in *Apple Inc.* should attach as **amod**. Nominal suffix designations that do not head the name, e.g. *LLC* ("limited liability corporation"), should attach as **nmod:desc**. For personal names, the suffix type *III* in (13b) is addressed above. Generational name suffixes that do not use numerals, like *Richard* ***Jr.*** and *Richard* ***the Third***, are treated as postmodifying **amod**. Other abbreviated name suffixes that would expand to nominal expressions, such as professional or honorary designations (*MD*, *O.B.E.*), attach as **nmod:desc**.

3.6 Nicknames and parenthetical descriptors

A nickname that takes the form of a full NP appended to a name, e.g. *Richard* ***the Lionheart***, can be attached as **appos**. The same goes for works of art featuring a formulaic name followed by a nickname: *Symphony No. 5 "****Fate****"*. Parenthetical descriptions following a name that are not alternate references to the entity should be treated as **parataxis**: *Pierre Vinken,* ***61****, said…*; *Vinken,* ***61 years old****, said…*; *The Chicago Manual of Style,* ***17th edition***; *Biden* ***(D)*** *said…* (but *Biden,* ***a Democrat****, said…* would be **appos**).

3.7 Addresses

A street address like *221b Baker St.* is headed by *St.*, with *Baker* attaching as **compound**, and *221b* per the policy on numbered entities (§3.4). Frequently, place descriptions specify a locale-NP postmodifier without a connective word besides punctuation. Examples: *London,* ***UK***; *University of Wisconsin–****Madison***; *CSI:* ***Miami***. These should be considered adverbial NPs, which arguably should fall under the **nmod:npmod** relation.[23]

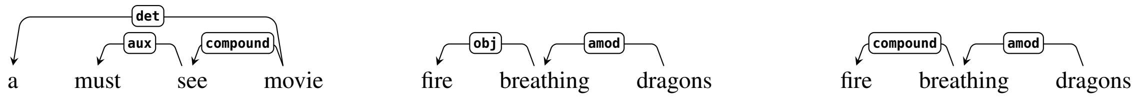
Multiple tokens of a single phone number should be joined with **flat** (this is the practice in the GUM corpus; EWT currently favors **nummod**). Separate pieces of metadata that are juxtaposed in an extralinguistic fashion (e.g., name, street address, city, postal code) should be treated as items of a list—successive items should attach to the first as **list**.

4 Phrasal Attributive Modifiers

In English, the attributive modifier position before the noun head in a noun phrase is not limited to adjectives/adjective phrases (*very easy to use*) and nominals. It also accommodates phrases like:

[22] Confirming native speaker intuitions, a search of COCA (Davies, 2010) reveals that the plural is much less frequent than the singular in the pattern N.PL *number(s)* NUM *and* NUM, with the exception of the abbreviated spelling, where *nos.* is more prevalent in this context than *no.* (the abbreviations seem to be especially conventional in proper names like *Symphony No. 5*).

[23] Currently, corpora sometimes use **nmod:npmod** and sometimes use **appos**, which is not appropriate as the two parts of the location are not interchangeable. Space does not permit full discussion of **nmod:npmod** here (but see Schneider and Zeldes, 2021, §6).



(a) Aux+V modifier **(b)** ‘Deep structure’ N+V modifier analysis **(c)** Proposed ‘surface structure’ analysis

Figure 2: Phrasal attributive modifiers (hyphen tokens omitted for brevity).

(15) a. a **high-quality** product
b. a **by-the-book** strategy
c. a **fly-by-night** operation
d. a **have-your-cake-and-eat-it-too** plan
e. a **come-to-Jesus**, **do-or-die** moment
f. a stern **don’t-mess-with-me** look
g. a **must-see** movie
h. **fire-breathing** dragons
i. the **Bible-thumping**, **church-going** faithful
j. many **so-called** libertarians
k. a **cost-effective**, **nuclear-free** future

Assuming that the hyphenated expressions are tokenized as separate words, UD annotators are confronted with two issues: how to analyze these phrases internally, and which dependency relation to use for the modification of the external noun.

Some of the hyphenated expressions in (15) are clearly lexicalized; others are productive combinations. Expressions of this type might loosely be described as ‘compounds’, in the sense that the joining of multiple content words into one lexical item is the morphological process of compounding. Should the hyphenated parts thus be joined together with **compound** across the board? We are hesitant to establish this policy because it would overload an already very broad relation label. Centrally, in noun phrases, **compound** describes modification of a noun by another noun. If it applies to the examples in (15), it would be for attachment to the underlined noun, not the internal structure of the hyphenated expression.

Another consideration is that the internal structure of the hyphenated phrases is *largely* regular: phrasal modifiers of nouns can be structured as modified nouns (15a), PPs (15b), VPs (15c, 15d), imperative sentences (15e, 15f), and verb clusters (15g). These structures are transparent, and just as UD policy analyzes regular internal structures in proper names like *University of Wisconsin*, we advocate recognizing internal structure here.

Yet synthetic or argument structure compounds such as *fire-breathing*, *Bible-thumping*, and *church-going* (15h, 15i) invert the normal clausal order. Neither *fire* nor *Bible* nor *church* is the subject in the clausal paraphrase: *fire* is the direct object in *breathing fire*; the paraphrase of *Bible-thumping* would require reordering and adding a determiner or plural for the direct object; and the paraphrase of *church-going* would require a preposition: *going **to** church*. Meanwhile, *so-called* (15j) lacks any obvious paraphrase as a clause. We take these anomalies in word order and morphosyntax as clear evidence that left-headed ‘deep structure’ VP material is being grafted onto a right-headed compound in the ‘surface structure’. As Basic UD aims to represent surface syntax, we join these expressions as **compound**, as shown for *fire-breathing* in figure 2c (vs. figure 2b). The adjective-headed combinations in (15k) should also use internal **compound**, as should numeric modifier compounds like *a **10-year** plan*.[24]

The next question is the external attachment, which is made difficult by UD’s lexicalist principle that the part of speech of a word determines which relations it can participate in. Consider *must-see* (15g), which is not a full VP, merely an auxiliary plus its head verb. Is this to be treated as a clausal dependent—**acl**, or even **acl:relcl** (a relative clause)? This seems dubious; note that a relative clause paraphrase would involve an embedded subject, e.g. *a movie that **one** must see*, or else a passive—*a movie that must be seen*. It is also doubtful whether (15c–15f) should be treated as clausal modification, yielding several different dependency labels for the attributive relationship. A simpler solution, it seems to us, is to treat attributive phrasal expressions internally headed by verbs like coerced noun phrases,[25] with **compound** for the external attachment, as shown in figure 2a. As for PP modifiers like in (15b), it seems simplest to attach

[24] Contrast *10-year* (**compound**) with *10 years* (**nummod**), where the number modifier controls agreement.

[25] Kahane et al. (2017) suggest expanding the UD notion of multiword token to include idiomatic phrasal expressions, separating their external syntactic behavior from their internal structure. This would make it convenient to represent the expression *must-see* as a multiword NOUN comprised internally of an AUX and a VERB. This could be indicated via a morphological feature ExtPos=NOUN on the internal head, *see*.

them as **compound** rather than **nmod**; on this view, English nominal **compound** is equivalent to attributive modification by a non-possessive nominal phrase (a hypothetical alternate name being **nmod:attr**).

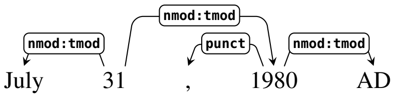
To summarize, our proposed policy for phrasal attributive modifiers of nouns is:

- The attributive expression is internally analyzed with regular relations to the extent possible, except where those relations defy ordinary word order or morphosyntax. **compound** is used internally for anomalous relations.
- In the interest of simplicity, all non-possessive attributive modifiers attach as either **compound** if internally headed by a nominal or nominalized phrase (including PPs), and **amod** etc. for adjectival heads, as appropriate.

5 Dates

While analytically expressed dates like *the thirty-first of July* follow normal syntax (with *thirty-first* elliptical for *thirty-first day*), there are special written formats for dates and times. Instead of a flat structure, which would obscure the compositionality of dates, we propose the simple principles of (a) treating the most precise part of the expression as its head, and (b) connecting the parts of the expression together with **nmod:tmod**.[26]

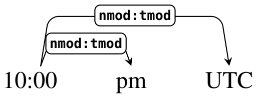
For example, *July 31, 1980 AD* consists of a year expression (*1980 AD*) and a month both modifying a date:



Another convention puts the date before the month (*31 July*). There, too, the date would be the head. Even when the date is written as an ordinal—*July the fourth*—the month should be considered a temporal modifier because it can be omitted with sufficient context (*I'll see you on the fourth*; **I'll see you on July*). This is in contrast to *Richard the Third* (§3.5), where *Richard* is the head.

A further practical consideration is that UD tree heads are often used to determine minimal token spans for annotations such as entity recognition, mentions in coreference resolution, and entity linking spans for Wikification (associating mentioned entities with their Wikipedia entries; Ratinov et al., 2011). Such minimal or 'MIN' spans (Poesio et al., 2018, p. 12) are then used for training and scoring systems in 'fuzzy' match scenarios. It makes intuitive sense for the day in date expressions to form the minimal span which needs to be identified, since the other tokens, i.e. years and months, already form the minimal spans for the nested mentions of those years and months as separate entities. This use of UD-tree heads is already in place for non-UD corpora using UD parses, such as ARRAU (Uryupina et al., 2020), and in the gold standard UD English GUM for NER, coreference and Wikification (Lin and Zeldes, 2021).

For time expressions we follow similar reasoning, with an example as follows:



The time zone could alternately be expressed as a phrase like *London time*, which we would also view as **nmod:tmod**. If written as *ten o'clock*, the token *o'clock* is considered an adverb and **advmod** of *ten*. This also corresponds to an etymological reading of *o'clock* (< *of clock*), since a univerbized prepositional phrase is equivalent to an adverb (cf. adverbs like *ashore*, formed with the Old English preposition *an*, the stressed equivalent of *on*).

Zeman (2021) likewise proposes a standard for dates and times (considering English as well as Czech, Indonesian, and Chinese). That approach is similar, differing mainly in treating the year in a date expression as headed by the month rather than the date—*1980* would be a dependent of *July*, which would be a dependent of *31*, in *July 31, 1980*. While semantically intuitive (smaller units of time head the next larger containing unit), it is not clear that there is any *syntactic* motivation to group the month and

[26] We considered finer-grained relations like **nmod:month**, **nmod:year**, **nmod:era**, **nmod:ampm**, and **nmod:tz** but concluded these were too detailed for UD and should fall under the purview of information extraction.

year together. Although the month cannot normally be omitted while retaining the year, an expression like *the 31st, 1980* is only semantically nonsensical, or at best pragmatically anomalous, but not truly ungrammatical. As evidence for this we consider the possibility of felicitous day+year expressions, such as *New Year's Day 2000* (the same as 2000-01-01) or *Pentecost 2022* (2022-06-05). The year-modifies-month approach also has the disadvantage of creating nonprojectivity if the date is written between the month and the year.

Zeman (§5) suggests **appos** to link a date with a day of the week, as in *Wednesday, July 31*. We agree with this policy. Though the day of the week conventionally comes first in English, we recognize that the order may be reversed on occasion (reversibility is a definitional criterion for **appos**, which is always left-headed). Moreover, this does not affect preposition choice, as *on* marks days of the week as well as dates, supporting the **appos** analysis in which they are essentially interchangeable full NPs.

6 How prevalent are these issues?

Some readers may wonder how common the issues raised in this paper actually are, and in particular whether their frequency merits adding relation subtypes such as **nmod:desc**. Table 2 gives statistics for some types of constructions that would be covered under the umbrella of such a relation. Although the phenomena are not extremely frequent, the total token count of 373 out of 152K tokens in the UD v2.9 edition of GUM puts a putative relation covering these at rank 35 of 49 relation labels (including subtypes), between **obl:tmod** (362 tokens) and **nmod:tmod** (399), suggesting that these are not particularly rare occurrences. We also presume that depending on genre, some subtypes may become much more frequent, such as company suffixes or even personal titles—for example, the frequency of just company suffixes in EWT seems is about 2.5 per 10K tokens, compared to 0.3 per 10K tokens in GUM (other categories are harder to identify, since their annotation in EWT currently varies or is not easily distinguishable, as in the case of numbering modifiers).

construction	**most frequent types**	**tokens (GUM)**	**types (GUM)**
title/profession	General (15), Mr. (10), St. (8)	202	78
numbering	Figure (31), Method (20), Wave (10)	162	63
company	Inc (4)	4	1
entity type	Mount (1), Camp (1), Team (1)	5	5
total		373	147

Table 2: Frequencies of some mischievous nominal constructions in GUM.

Although adding a new labeling distinction in the form of **nmod:desc** would doubtless require some manual disambiguation effort, we feel that by surveying the constructions in this paper in detail, it becomes more feasible to design high recall, automatic approaches to creating an initial updated version of UD English with a more nuanced treatment of these mischievous constructions, using UD editing libraries such as DepEdit (Peng and Zeldes, 2018) or Udapi (Popel et al., 2017), which can then be subjected to a manual filtering pass.

7 Conclusion

Above we have reviewed many constructions involving names, values, and compounds that have pointed to blind spots in the current guidelines for the **nmod:***, **compound**, **flat**, **appos**, and **nummod** relations. We have laid out several options for improving the treatment of these constructions via clearer and more principled guidelines. The proposed improvements are of a surgical nature, minimizing disruption to other UD conventions (no new universal relations are proposed, for instance). We are cognizant that considerable effort may be required to fully revise existing UD treebanks, but note that treebanks are already inconsistent; clearer guidance can only help. Subtypes remain officially optional—it is not necessary for a treebank to distinguish subtypes of **nmod** to be compliant with the UD standard.

We invite feedback on these proposals from the UD community, particularly with regard to other languages. We are aware that treebanking efforts in other languages have encountered some of the same issues, but we have not systematically investigated our proposed solutions beyond English.

Acknowledgments

Though the synthesis of problems, argumentation, and recommendations presented above are new, many of the specific challenges and alternatives arose from background discussions on GitHub about the UD guidelines. These discussions took place over several years in more than a dozen discussion threads involving (at least): Aryaman Arora, Colin Batchelor, Flavio Massimiliano Cecchini, Xinying Chen, Çağrı Çöltekin, Sylvain Kahane, Chris Manning, Bohdan Moskalevskyi, Joakim Nivre, Martin Popel, Alexandre Rademaker, Livy Real, Jack Rueter, Sebastian Schuster, Francis Tyers, Jonathan North Washington, and Dan Zeman. We are grateful to all who provided insights.

References

Shabnam Behzad and Amir Zeldes. 2020. A cross-genre ensemble approach to robust Reddit part of speech tagging. In *Proc. of the 12th Web as Corpus Workshop*, pages 50–56, Marseille, France.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4):447–464.

Nanna Fuhrhop. 1996. Fugenelemente. In Ewald Lang and Gisela Zifonun, editors, *Deutsch - typologisch*, pages 525–550. de Gruyter, Berlin.

Rodney Huddleston and Geoffrey K. Pullum, editors. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - Propositions for Universal Dependencies. In *Proc. of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 181–189, Prague, Czech Republic.

Jessica Lin and Amir Zeldes. 2021. WikiGUM: Exhaustive entity linking for Wikification in 12 genres. In *Proc. of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 170–175, Punta Cana, Dominican Republic.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proc. of LREC*, pages 4027–4036, Marseille, France.

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, and Daniel Zeman. 2016. Universal Dependencies v1: a multilingual treebank collection. In *Proc. of LREC*, pages 1659–1666, Portorož, Slovenia.

Siyao Peng and Amir Zeldes. 2018. All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proc. of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 167–177, Santa Fe, New Mexico, USA.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proc. of the First Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2018)*, pages 11–22, New Orleans, LA.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proc. of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proc. of ACL-HLT*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. FrameNet II: extended theory and practice.

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. *arXiv:2108.12928 [cs]*.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proc. of LREC*, pages 2897–2904, Reykjavík, Iceland.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus. *Natural Language Engineering*, 26:95–128.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency Treebank. In *Proc. of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 192–201, Brussels, Belgium.

Daniel Zeman. 2021. Date and time in Universal Dependencies. In *Proc. of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*.