Adpositions in Universal Dependencies

Leaves

Adpositions (i.e. prepositions and postpositions) should be attached as leaves. One known exception is multi-word adpositions where the non-first words are attached to the first one as `mwe`, even if the first word is preposition. Also the whole MWE may act as something else than adposition; for instance, it can be an adverbial modifier, hence it will be labeled `advmod`.

There are numerous cases in UD 1.1 where this condition does not hold. To search for them, go to PML-TQ and select one of the Universal Dependencies treebanks, e.g.

http://lindat.mff.cuni.cz/services/pmltq/ud_de/

for German. Then enter the following query:

```
a-node [tag="ADP", child a-node [deprel!="mwe"]]
```

Case vs. mark

An adposition modifying a noun, pronoun, adjective, determiner or numeral should normally be labeled `case`. If it modifies a verb, it should be labeled `mark`. Thus it is not true that all prepositions are `case` and all subordinating conjunctions are `mark`. At least in some languages prepositions (adpositions) can be `mark` as well. Moreover, a subordinate clause may be headed by a non-verb if there is a nominal predicate (with or without copula), a passive or an ellipsis; in these cases an adposition or subordinating conjunction will be labeled `mark` even though it is not attached to a verb (example: *He could speak two languages before he was two years old.* … mark(old, before)).

The UD definition: "A marker is the word introducing a finite clause subordinate to another clause." I am not sure to what extent the restriction to *finite* clauses is intentional here; even the universal description page contains a German example where the clause is in infinitive and both the subordinating conjunction / preposition *um* and the infinitive marker / preposition *zu* are labeled `mark`. (*Er kam wieder, um das Werk zu Ende zu bringen.* "He came again to bring the work to end.") And the corresponding English documentation explicitly says that "the infinitive marker *to* is analyzed as a `mark`."

There are languages / treebanks where prepositions are labeled `case` even when they are attached to verbs. Try the following query in PML-TQ to find them:

```
a-node [tag="ADP", deprel="case", parent a-node [tag="VERB"]]
```

In some cases these are probably just treebank conversion errors. In others, they seem to be systematic and perhaps intentional? For example the preposition-infinitive relations in Spanish are frequently (always?) labeled `case`; maybe this is a consequence of the word *finite* accidentally introduced in the universal definition? Example: *empresas interesadas en prestar el servicio* "companies interested in providing the service". There is a comparable example in the English data: … *New Delhi does better in providing security* … and in this case, the relation between *in* and

providing is labeled `mark`. The above query yields only one result in the English UD 1.1, and that seems to be a problem of the POS tagging, not of the dependency analysis (*crossing* is tagged as `VERB` but it should be a `NOUN`).

The `case(infinitive, preposition)` relation appears in several other languages. I found a few interesting occurrences in Italian where there is also a definite article before the infinitive *(da il negare l' esistenza),* which might suggest that the infinitive is actually much more closer to a noun here, even though it is still tagged `VERB`. There is an analogy in Uralic languages: Finnish *tervehtimässä* "greet" is an infinitive with the inessive case suffix. We could say that this supports the view that prepositions with infinitives can actually be labeled `case`. But where is the borderline between `case` and `mark` then?

BTW, infinitives with articles are less frequent than those without articles in Italian. The following query gave only 47 results:

```
a-node [tag="VERB", iset/verbform="inf", child a-node [tag="ADP",
deprel="case"], 1+x child a-node [deprel="det"]]
```

The following query gave 84 results:

```
a-node [tag="VERB", iset/verbform="inf", child a-node [tag="ADP",
deprel="case"], 0x child a-node [deprel="det"]]
```

Mark vs. aux

Let's modify one of the previous queries so that we see other non-mark relations, not only case.

```
a-node [tag="ADP", deprel!="mark", parent a-node [tag="VERB"]]
```

In English, this will now return hundreds of other uses, e.g. `compound:prt` or promotions to the place of a missing noun *(for outcomes that they had little to do with* … nmod(do, with)).

But let's try Spanish. Here we now see that the preposition *a* is not even labeled `case`, it is `aux`. One could argue that it is used with infinitives in periphrastic verb forms *(volvió a aumentar* lit. *returned to increase* "increased again"; *volvió* is also labeled `aux`). But shouldn't the `aux` relation be reserved for auxiliary *verbs?* Indeed the UD definition thinks so: "An auxiliary of a clause is a non-main verb of the clause, e.g., a modal auxiliary, or a form of *be, do* or *have* in a periphrastic tense."

If we further extend this investigation to non-verbs other than adpositions, we will find in Bulgarian that the particles *да / da* and *ще / šte* are also labeled `aux`. Shouldn't they be `mark` instead? This is the query:

```
a-node [tag!~"AUX|VERB", deprel="aux", parent a-node [tag="VERB"]]
```

Some of the search results:

Можеш ли да плуваш? / Možeš li da pluvaš? "Can you swim?"
Който е богат, ще плаща повече. / Kojto e bogat, šte plašta poveče. "Who is rich will pay more."

I believe that the particle *да* is somewhat similar to the complementizing subordinating conjunctions in other (not only Slavic) languages, and also somewhat similar to the infinitive

markers in Germanic languages (because there is no morphological infinitive in Bulgarian). Both subordinating conjunctions and infinitive markers are labeled `mark` in other UD datasets.

Greek: the particles *να / na*, *ας / as* and *θα / tha* are attached as `aux`. But this time it is clearly stated in the documentation of Greek. The particles are used to form subjunctive and future, respectively. (Seems like an analogy to Bulgarian, so we probably want to solve both languages the same way.)

In Basque, a number of `SCONJ` are attached as `aux`.

German: the infinitive marker *zu* seems to be always labeled `aux`. Other prepositions (like *um* in *um zu* + infinitive) are also labeled `aux`. (Moreover, the POS tag of *zu* is `PART` but I believe it should be `ADP`.)

Spanish: 232 occurrences of the preposition *a*, 22× *de*, and a few others.

In English, there are 6 results, most of them featuring the adverb *better* (maybe because it functions similarly to the modal *should*? But I don't see any mention of this in the English UD documentation.)

Croatian, French, Indonesian and Italian: only a few hits each, probably annotation or transformation errors.

Compound:prt vs. mark

In Germanic languages, some adpositions may be used as verbal particles (English *on* in *come on*) or separable verb prefixes (German *aus* in *aussehen => er sah aus*). These words keep the POS tag of adposition (`ADP`; not `PART`) but their relation to the main verb node is a subtype of `compound`, to reflect that the verb + the particle (adposition) actually form a new lexical unit with different meaning. The subtype is called `compound:prt` and is used in English, Swedish and Danish so far.

For some reason (is there a reason?) it is not used in German, where the separated verb prefixes are attached as `mark` instead.