Towards an adequate account of parataxis in Universal Dependencies

Lars Ahrenberg
Department of Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Abstract

The parataxis relation as defined for Universal Dependencies 2.0 is general and, for this reason, sometimes hard to distinguish from competing analyses, such as coordination, conj, or apposition, appos. The specific subtypes that are listed for parataxis are also quite different in character. In this study we first show that the actual practice by UD-annotators is varied, using the parallel UD (PUD-) treebanks as data. We then review the current definitions and guidelines and suggest improvements.

1 Introduction

The aims of the Universal Dependencies (UD) project are high and somewhat conflicting, (cf. Osborne and Gerdes, 2019). While the emphasis is on linguistic typology and parsing, there are no restrictions on the kind of linguistic data that UD could be applied to. Indeed, the UD treebanks cover as varied genres as tweets and literature. Adding to this the desired goal that UD parsers should have high accuracy on text which has not been tokenized, we have a situation where UD has to deal not just with common syntactic constructions, but with a multitude of genre-specific and stylistic varieties.

As an annotator one frequently encounters cases where the choice between *parataxis* and some other relation is unclear. Take (1) as an example:[1]

(1) A very good performer she was; breaking into Yiddish, into Italian, into German, accenting and gesturing, turning now into a claque of elderly Jews, now into a frightened small boy.

We have one typical instance of parataxis, supported by the semi-colon and relating the head of what follows it, *breaking*, to the head of the clause preceding it, *performer*. But what about the verbs *accenting* and *turning*? Are they conjuncts to *breaking* or related via parataxis or maybe adverbials? The sequence of prepositional phrases, *into Yiddish, into Italian, into German* could similarly be seen as coordinated with *Yiddish* as the head, related via parataxis, or even be analysed as independent oblique constituents in relation to *breaking*. The last part of the sentence forms what Matthews (**?**) calls a correlative construction; two clauses or phrases held together by occurrences of a word pair, here *now – now*, the relation of which is open to interpretation.

Part of the problem for a UD annotator here is that UD does not require a conjunction to be present with a coordinated conjunct. As a matter fact, the classical quote attributed to Caesar, *Veni, vidi, vici*, is analysed as a coordination in the UD guidelines[2], while in many works in stylistics, it tends to be described as a prime example of parataxis[3]. Another factor is that UD so far has not been specific enough about the categories that can enter into a parataxis relation.

The purpose of this paper is twofold. First, we study how the parataxis relations and its competitors have been used in the Parallel UD (PUD) Treebanks. The choice of these treebanks is motivated by the fact that the sentences in them are close translations of one another, mostly from English source

[1] All examples that are not from the PUD treebanks are taken from, or modelled upon, sentences found in (**?**)
[2] https://universaldependencies.org/u/dep/conj.html
[3] See, for example, https://literarydevices.net/parataxis/

sentences. Thus, their annotations should be as similar as can be found in the full set of UD treebanks. Second, in view of problems we have encountered in the annotation of paratactic style in literature, we review the current guidelines for paratactic relations, in particular *parataxis*, with the goal of suggesting improvements and clarifications.

Paratactic relations are of interest both to grammar and stylistics. Matthews (**?**) argues that there is no clear-cut border, and this is, we believe, partly what causes the difficulty for UD annotators. Halliday (**?**) provides a detailed analysis of the opposition hypotaxis-parataxis within his broader systemic model that also involves 'logico-functional' aspects of relations between syntactic units. This level is however not available in UD representations.

2 Paratactic relations in UD

The UD framework offers a limited number of paratactic relations, or loose joining relations as they are called in (de Marneffe et al., 2014). The most common are *conj* that covers all forms of coordination, *parataxis* used for (mostly) asyndetic sequences of clauses, and *appos* for co-referring nominals that come one after the other, with or without a comma in between. Other paratactical relations are *list*, used for sequences of small information units, *flat* used for multi-token names, *fixed* for fixed lexical items, and *discourse* for interjectional items.

The relations *list*, *parataxis*, and *appos* are needed as they give 'a robust analysis of more informal forms of text (de Marneffe et al., 2013; de Marneffe et al., 2014). Parataxis is said to be needed also with more formal writing for constructions such as sentences joined with a colon. In (de Marneffe et al., 2013) we get an example of the use of list: *Steve Jones Phone:555-9814 Email:jones@abc.edf* where *Phone* and *Email* depends on *Jones* via *list*. This example is found also in the guidelines, though *Steve* is now the head node.

In the current guidelines, *parataxis* is described as 'a relation between a word (often the main predicate of a sentence) and other elements, such as a sentential parenthetical or a clause after a ':' or a ';', placed side by side without any explicit coordination, subordination, or argument relation with the head word. The guidelines go on to list five sub-types: side-by-side sentences, reported speech, interjected clauses, tag questions, and news article bylines.

The guidelines for other languages follow the ones for English to a large extent. Some languages, including Italian (**?**) and French (Gerdes and Kahane, 2017), have developed language-specific extensions for parataxis.

Apart from the other paratactic relations, we find several hypotactic relations, and especially *ccomp*, that competes with *parataxis* for the analysis of certain constructions.

3 Parataxis in the PUD treebanks

Most of the parallel UD treebanks were developed for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (**?**). Others have been developed afterwords, such as those for Czech, Finnish, and Swedish. All PUD treebanks contain 1000 sentences of which the first 750 are originally English, while the remaining are originally German, French, Italian, or Spanish.

Translations to other languages were made via English. For this reason, we take the English PUD treebank as our reference. It has 97 instances of parataxis, the majority of which are either side-by-side sentences or reported speech. The rest we have divided into Parentheticals and Other (see Table 1). Parenthetical covers interjected clauses and news article bylines, but the majority are inserted or added

Subtype	**Frequency**
Side-by-side sentences	44
Reported speech	42
Parentheticals	8
Tag questions	0
Other	3
Total	97

Table 1: The parataxis relation in English PUD distributed on subtypes.

comments of the kind illustrated in (2).[4] The Other class is used for a few cases which we believe are better analysed as something else. An example is (3).

(2) (a) Their first king was **Mojmír** I (**ruled** 830–846).
(w01010047) *parataxis(Mojmir,ruled)*
(b) And, she granted, "you have to look at where she has acknowledged that we **need** to do something different–we can **do** better–and where she has expressed regret."
(n01060069) *parataxis(need,do)*

(3) Where does all her energy **come** from? Or that **voice**, which can blast out with a force to induce shockwaves? (n01116018) *parataxis(come,voice)*

The translators were asked to produce as close translations as possible (**?**). Thus, the translations can be expected to follow the structure of the source sentences. This is largely the case, but some languages, such as German, have more constructional differences than others with English. Table 2 shows absolute frequencies for the parataxis relation in the 14 PUD treebanks that have more than 10 instances of it. The variation in numbers is striking given that content and structure are supposedly very similar. The differences become even more pronunced when we look at the distribution of the parataxis relation in the treebanks. Table 3 shows the number of overlapping and non-overlapping relations for English compared to the other languages.

Treebank	**en**	**ar**	**cs**	**de**	**es**	**fi**	**fr**	**hi**	**id**	**it**	**pt**	**ru**	**sv**	**tu**
Frequency	97	24	20	68	106	108	106	93	116	99	103	195	134	74

Table 2: Absolute frequencies for the *parataxis* relation in fourteen PUD treebanks: English, Arabic, Czech, German, Spanish, Finnish, French, Hindi, Indonesian, Italian, Portuguese, Russian, Swedish, and Turkish.

Treebank	**Overlaps**	**English only**	**Other only**	**Similarity**
UD_Arabic-PUD	15	82	9	0.24
UD_Czech-PUD	10	87	10	0.17
UD_Finnish-PUD	73	24	35	0.71
UD_French-PUD	62	35	42	0.62
UD_German-PUD	43	54	25	0.53
UD_Hindi-PUD	40	57	53	0.42
UD_Indonesian-PUD	44	53	70	0.43
UD_Italian-PUD	73	24	26	0.75
UD_Portuguese-PUD	79	18	24	0.79
UD_Russian-PUD	74	23	112	0.52
UD_Spanish-PUD	72	25	34	0.71
UD_Swedish-PUD	86	11	48	0.75
UD_Turkish-PUD	43	54	26	0.52

Table 3: How English instances of *parataxis* are distributed compared to instances in other PUD treebanks. Overlaps have been assumed whenever an English tree and the corresponding tree for the other language have the same number of instances. Similarity is measured as $2 * Overlap/(English + Other + 2 * Overlap)$

Using a simple similarity metric, we can see that the Portuguese treebank is the most similar in its annotation of the parataxis relation. Still, even the English-Portuguese pair has more than 40 cases of non-overlaps. The Swedish treebank has transfered the highest share of English instances (86 out of 97) but has added many extra ones where the English PUD uses hypotactic relations.

A part of the variation can be explained by constructional changes made in the translations, but the found variation is much to great to suggest that the goal of "consistent annotation of similar constructions across languages"[5] has been met.

[4]The label enclosed in parenthesis is the sentence identifier, sent_id, from the PUD files.

[5]https://universaldependencies.org/introduction.html

4 Problematic constructions

A closer look at the PUD data reveals a number of typical cases, where the annotation between languages differ, although the constructions are similar.

4.1 Side-by-side clauses

Asyndetic side-by-side clauses can be annotated either as *parataxis* or *conj*. Both interpretations accord with the guidelines (cf. the Latin example in the introduction). As the guidelines don't provide distinctive information, the decision is up to the annotator's judgement. The following is an English-Italian pair.

(4) EN: I'm **going** to jail either way, **hope** it was worth it
(n01011017) *parataxis(going,hope)*
IT: In entrambi i casi **finirò** in prigione, **spero** ne sia valsa la pena
(n01011017) *conj(finirò,spero)*

German, Swedish and Spanish have also opted for parataxis, whereas French follows Italian, using conj. It should be noted that the presence of a coordinating conjunction is not always disambiguating for annotators. Sentence (3) above is a telling example.

4.2 Reported speech

The guidelines for reported speech clearly distinguish clausal complements from paratactical direct speech. Basically, the difference is that when the reported speech is, or could be, introduced by a subjunction such as *that*, the analysis should be *ccomp*, otherwise it should be *parataxis*. In particular, if there is a clear signal of separation, such as a colon or comma and/or citation marks, *parataxis* should be used.

Many annotators, though, prefer to see the speech verb as governor and the reported speech as complement. This is implemented throughout in the German PUD treebank. Here both the relation and the direction of the relation changes. You may also find examples where the parataxis relation is used, but reversed, as in the French version of (5).

(5) EN: "This is a **disaster** for pain patients," Mailis **said** in an interview ...
(n01041006) *parataxis(disaster,said)*
DE: Das ist eine **Katastrophe** für Schmerzpatienten, **sagte** Mailis in einem Interview ...
(n01041006) *ccomp(sagte,Katastrophe)*
FR: C'est un **désastre** pour les malades en souffrance, a **déclaré** Mailis dans un entretien ...
(n01041006) *parataxis(déclaré,désastre)*

4.3 Parentheticals

In the category of Parentheticals we find cases where parataxis alternates with appos. In the following example, the inserted unit is a noun phrase that may be interpreted as an elliptical clause. English PUD annotates this as parataxis, while Spanish (and German, Swedish) treats it as apposition. The Spanish translation has placed the verb translating *sent* after the phrase referring to dinosaurs. This may explain the actual annotation, but should not really affect it. In French, the translation has inserted the conjunction *et*, suggesting an analysis as conjunct. There is no conjunction in the English and Spanish sentences, but it may nevertheless be inserted (or 'heard') without changing the meaning much.

(6) EN: ... and sent so many **species** - not just the **dinosaurs** - into oblivion .
(n01023034) *parataxis(species,dinosaurs)*
ES: ... y que hizo que muchas **especies**, no solo los **dinosaurios**, cayeran en el olvido.
(n01023034) *appos(species,dinosaurs)*
FR: ... qui ont causé l'extinction de nombreuses **espèces**, et pas uniquement des **dinosaures**.
(n01023034) *conj(espèces,dinosaures)*

4.4 Hypotactic competitors

Apart from the differences in the analysis of reported speech, there are other cases where a parataxis relation in one language corresponds to a hypotactic dependency in another language without any apparent change of structure. The Swedish PUD treebank has several examples of parataxis, where English and other languages prefer *obl*, as in (7). A possible reason for this analysis is the presence of a comma, indicating a pause, before the prepositional phrase.

(7) EN: The issuing of coinage is predominantly **numismatic** in nature, with the **intention** of being sold mainly to collectors. (w04003054) *obl(numismatic,intention)*
SV: Utfärdandet av mynt är företrädesvis **numismatiskt** till sin natur, med **avsikten** att säljas främst till samlare. (w04003054) *parataxis(numismatiskt,avsikten)*

5 Discussion

What is to be done? We could put some of the blame on annotators who don't follow the guidelines properly. But annotators have intuitions, often based in a linguistic tradition, and may violate the guidelines for good reasons.

We could also blame the guidelines for not being complete or clear enough. This is partly true but the guidelines for reported speech are very clear, providing ample examples to clarify the contrast of parataxis to ccomp. But the frequency of ccomp in the analysis of reported speech raises the question why this should be a case of parataxis in the first place. Reported speech is the only type of parataxis where one of the two units is a semantic argument of the other and thus it is motivated to treat it as a case of complementation. The distinction between direct and indirect speech can be done as a specialization of ccomp, if needed.

An even stronger recommendation would be to generally prefer hypotactic relations over paratactic ones in UD, when there are arguments for both alternatives. This would make it clear that a sentence such as (7) has no parataxis relation.

Conversely, we note that UD has authorities such as Halliday (**?**) behind it in treating reported speech as a case of parataxis. It can also be argued that such a move would make the analysis of sentences such as (8) more complicated, as the clausal argument of the speech verb would then be split causing the relation between the two parts to cross the root node. However, non-projectivity cannot be entirely avoided with the current recommendation either and, actually the tree also relates 'leave' to 'time' as the head of a relative clause. The interposing of units is quite common in literary genres and, as they are easy to detect, could actually be given a dependency relation of their own, which as other relations may be further subtyped if needed. To compensate for this addition to the framework, the relation *list* could be deprecated as the kind of constructions it has been used for could equally well be modeled with parataxis.

(8) There **was** a time, Mr Panvalkar **said**, when he felt that they should leave the building.
(n01010042) *parataxis(was, said)*

Turning to the side-by-side sentences, there are two oppositions where annotations typically differ: parataxis vs. conjunct, and parataxis vs. apposition. For the first pair one could take the view that the presence of a coordinating conjunction somewhere in the sequence of units should always result in a conjunct analysis. This is a formalistic approach, but the alternative leaves much to annotators' varying intuitions. Moreover, when no coordinating conjunction is present, we need to know what information can guide an analysis as conjunct. It is possible to test whether a conjunction can be inserted without change of meaning and use this as a guideline. Sometimes, however, the insertion of a conjunction would lead to a substantial change of style, and thus be felt as the use of a different construction. This happens, inter alia, when a unit is repeated or slightly varied for reasons of emphasis or other stylistic effect, as in (9).

(9) (a) It's my mistake, my mistake.
(b) I had to let go of my detachment, my resentment.

Although the two units are filling the same slot in the larger sentential context, they are not joined by a coordinating conjunction, and cannot be without changing the construction's character. Given that the units are noun phrases, and in some sense, co-referring, it is tempting to analyse their relation as appositional. However, clauses can also be similarly repeated or varied, and this may speak in favour for an analysis as parataxis.

Another type repeats a structure, but each unit adds a different aspect, as in (10). In this case it is easy to hear a 'but' before the last unit in the sequence and treat the whole as a sequence of conjuncts. On the other hand, if the author had wanted a conjunction there, she would presumably have put one in. It would be somewhat disrespectful to ignore her choice.

(10) Her waist was curved, her legs were long, her breasts round, her stomach was flat, her bottom was not.

As regards parataxis vs. apposition the guidelines for the appos relation says that it relates a noun (head of a noun phrase) to a nominal, where the latter, dependent part is often optional. If parataxis relates clauses to one another, and appos relate nominals, we must nevertheless reckon with cases where a clausal head has a nominal side-by-side dependent, and this often leads to different analyses, as in (7) above. Moreover, we have sentences where the choice of head is not evident. Witness (11):

(11) EN: Greece was **divided** into many small, self-governing communities, a **pattern** largely dictated by Greek geography: ...
(w03005015) *parataxis(divided, pattern)*
DE: Griechenland war in viele kleine eigenständige Kommunen **unterteilt** - eine **Form**, die weitgehend durch die griechische Geografie vorgegeben ist: ...
(w03005015) *obl(unterteilt, Form)*
FR: Le pays est alors divisé en une **multitude** de petites communautés indépendantes, **situation** imposée par la géographie grecque
(w03005015) *appos(multitude, situation)*

In the French annotation, a noun is selected as head and the relation is taken as an apposition. In German the annotators have opted for an oblique relation which might be an error. Italian, Spanish and Swedish follow the English analysis. In the reverse situation, i.e., with a nominal head and a clausal dependent, the recommended analysis is acl, even for non-restrictive relative clauses.

(12) By comparison, it cost $103.7 million to build the NoMa infill Metro **station**, which **opened** in 2004. (n01005023) *acl:relcl(station,opened)*

However, adverbial and prepositional phrases can also be sequenced and then the distinction between hypotaxis and parataxis gets blurred. Look again at (1), repeated here for convenience as (13):

(13) A very good performer she was; breaking into Yiddish, into Italian, into German, accenting and gesturing, turning now into a claque of elderly Jews, now into a frightened small boy.

The *obl* relation has in principle no limitation on its number of occurrences in a single clause. For this reason one can be motivated to annotate not only *Yiddish* but also *Italian* and *German* as dependent on *breaking* via obl. A similar analysis could be proposed for *boy* with *turning* as its head. However, oblique units that belong to the same slot of the predicate have a stronger affinity than those who do not, which motivates a paratactical analysis. And although there are several places where the conjunction *and* can be inserted, suggesting a conjunct analysis, its absence may be taken as a decisive criterion for using parataxis.

(14) I could see her, hair and flesh escaping, hope trapped inside.

A similar problem arises in (14): should *escaping* be seen as dependent on *see* or *her*? In the first case we can choose between parataxis and advcl (hearing an absent *with* before *hair*), in the second acl would be the most appropriate relation. In a case like this, one cannot escape relying on annotators' individual interpretations, but in many of the other examples we believe a further elaboration of the guidelines could be quite helpful. Suggestions are given in the summary:

- The use of the *parataxis* relation varies quite a bit among UD annotators. This is not only caused by language or data differences, but differences in annotation practices, as evidenced by our study of annotations in the PUD treebanks.
- There is no need for more than one general and broadly defined relation of paratactic sequencing, *parataxis*. This relation may be further subtyped for the genres and treebanks that need it, as done for UD_Italian-Postwita (**?**) and UD_French-Spoken (Gerdes and Kahane, 2017). The *list* relation can be deprecated.
- The guidelines on *parataxis* need to be developed. The account of reported speech sets a good standard in terms of the level of detail, but should be motivated further.
- Just as UD have other basic principles, such as favouring content words and left constituents as heads, a basic principle stating preferences for hypotactic relations over paratactic could be adopted. Exceptions can be allowed, as is currently done for reported speech, and, as suggested here, for asyndetic sequences of units (clauses and phrases) expressing the same semantic role.
- Detailed guidelines regarding the possibility to determine a relation on the basis of inserting words that are not there should be worked out. The simplest guideline would be to completely disallow such argumentation except in cases of ellipsis, when the missing word can be inferred from the immediate context.

References

Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Samuel R. Bowman, Timothy Dozat and Christopher D. Manning. 2013. More constructions, more genres: Extending Stanford dependencies. *Proceedings of the 13th International Conference on Dependency Linguistics.*

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* 4585–4592.

Kim Gerdes and Sylvain Kahane. 2017. Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral: le cas de la macrosyntaxe. *Actes de la 24e conférence sur le traitement automatique des langues (TALN), Atelier sur les corpus annotés du français (ACor4French), Orléans.*

M. A. K. Halliday 1985. *An Introduction to Functional Grammar*. Edward Arnold.

P. H. Matthews 1981. *Syntax*. Cambridge University Press.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics* 4(1): 17. 1–28. DOI: https://doi.org/10.5334/gigl.537.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli and Fabio Tamburini. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).* https://www.aclweb.org/anthology/L18–1279.

Jennette Winterson 1997. *Gut symmetries.* Granta Books, London.

Daniel Zeman, Martin Popel, Milan Straka, ..., and Josie Li 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, 3–4 August 2017.* Association for Computational Linguistics.