Expletives in Universal Dependency Treebanks

Gosse Bouma[*°] **Jan Hajic**[†°] **Dag Haug**[‡°] **Joakim Nivre**[•°] **Per Erik Solberg**[‡°] **Lilja Øvrelid**[⋆°]

[*]University of Groningen, Centre for Language and Cognition
[†]Charles University in Prague, Faculty of Mathematics and Physics, UFAL
[‡]University of Oslo, Department of Philosophy, Classics, History of Arts and Ideas
[•]Uppsala University, Department of Linguistics and Philology
[⋆]University of Oslo, Department of Informatics
[°]Center for Advanced Study at the Norwegian Academy of Science and Letters

Abstract

Although treebanks annotated according to the guidelines of Universal Dependencies (UD) now exist for many languages, the goal of annotating the same phenomena in a cross-linguistically consistent fashion is not always met. In this paper, we investigate one phenomenon where we believe such consistency is lacking, namely expletive elements. Such elements occupy a position that is structurally associated with a core argument (or sometimes an oblique dependent), yet are non-referential and semantically void. Many UD treebanks identify at least some elements as expletive, but the range of phenomena differs between treebanks, even for closely related languages, and sometimes even for different treebanks for the same language. In this paper, we present criteria for identifying expletives that are applicable across languages and compatible with the goals of UD, give an overview of expletives as found in current UD treebanks, and present recommendations for the annotation of expletives so that more consistent annotation can be achieved in future releases.

1 Introduction

Universal Dependencies (UD) is a framework for morphosyntactic annotation that aims to provide useful information for downstream NLP applications in a cross-linguistically consistent fashion (Nivre, 2015; Nivre et al., 2016). Many such applications require an analysis of referring expressions. In co-reference resolution, for example, it is important to be able to separate anaphoric uses of pronouns such as *it* from non-referential uses (Boyd et al., 2005; Evans, 2001; Uryupina et al., 2016). Accurate translation of pronouns is another challenging problem, sometimes relying on co-reference resolution, and where one of the choices is to not translate a pronoun at all. The latter situation occurs for instance when translating from a language that has expletives into a language that does not use expletives (Hardmeier et al., 2015; Werlen and Popescu-Belis, 2017). The ParCor co-reference corpus (Guillou et al., 2014) distinguishes between anaphoric, event referential, and pleonastic use of the English pronoun *it*. Loáiciga et al. (2017) train a classifier to predict the different uses of *it* in English using among others syntactic information obtained from an automatic parse of the corpus. Being able to distinguish referential from non-referential noun phrases is potentially important also for tasks like question answering and information extraction.

Applications like these motivate consistent and explicit annotation of expletive elements in treebanks and the UD annotation scheme introduces a dedicated dependency relation (`expl`) to account for these. However, the current UD guidelines are not specific enough to allow expletive elements to be identified systematically in different languages, and the use of the `expl` relation varies considerably both across languages and between different treebanks for the same language. For instance, the manually annotated English treebank uses the `expl` relation for a wide range of constructions, including clausal extraposition, weather verbs, existential *there*, and some idiomatic expressions. By contrast, Dutch, a language in which all these phenomena occur as well, uses `expl` only for extraposed clausal arguments. In this paper, we provide a more precise characterization of the notion of expletives for the purpose of UD treebank annotation, survey the annotation of expletives in existing UD treebanks, and make recommendations to improve consistency in future releases.

2 What is an Expletive?

The UD initiative aims to provide a syntactic annotation scheme that can be applied cross-linguistically, and that can be used to drive semantic interpretation. At the clause level, it distinguishes between core arguments and oblique dependents of the verb, with core arguments being limited to subjects (nominal and clausal), objects (direct and indirect), and clausal complements (open and closed). Expletives are of interest here, as a consistent distinction between expletives and regular core arguments is important for semantic interpretation but non-trivial to achieve across languages and constructions.

The UD documentation currently states that `expl` is to be used for *expletive or pleonastic nominals, that appear in an argument position of a predicate but which do not themselves satisfy any of the semantic roles of the predicate*. As examples, it mentions English *it* and *there* as used in clausal extrapostion and existential constructions, cases of true clitic doubling in Greek and Bulgarian, and inherent reflexives. Silveira (2016) characterizes `expl` as *a wildcard for any element that has the morphosyntactic properties associated with a particular grammatical function but does not receive a semantic role*.

It is problematic that the UD definition relies on the concept of argument, since UD otherwise abandons the argument/adjunct distinction in favor of the core/oblique distinction. Silveira's account avoids this problem by instead referring to grammatical functions, thus also catering for cases like:

(1) He will see to *it* that you have a reservation.

However, both definitions appear to be too wide, in that they do not impose any restrictions on the form of the expletive, or require it to be non-referential. It could therefore be argued that the subject of a raising verb, like *Sue* in *Sue appears to be nice*, satisfies the conditions of the definition, since it is a nominal in subject position that does not satisfy a semantic role of the predicate *appear*.

It seems useful, then, to look for a better definition of expletive. Much of the literature in theoretical linguistics is either restricted to specific languages or language families (Platzack, 1987; Bennis, 2010; Cardinaletti, 1997) or to specific constructions (Vikner, 1995; Hazout, 2004). A theory-neutral and general definition can be found in Postal and Pullum (1988):

> [T]hey are (i) morphologically identical to pro-forms (in English, two relevant forms are *it*, identical to the third person neuter pronoun, and *there*, identical to the non-proximate locative pro-adverb), (ii) nonreferential (neither anaphoric/cataphoric nor exophoric), and (iii) devoid of any but a vacuous semantic role. As a tentative definition of expletives, we can characterize them as pro-forms (typically third person pronouns or locative pro-adverbs) that occur in core argument positions but are non-referential (and therefore not assigned a semantic role).

Like the UD definition, Postal and Pullum (1988) emphasize the vacuous semantics of expletives, but understand this not just as the lack of semantic role (iii) but also more generally as the absence of reference (ii). Arguably, (ii) entails (iii) and could seem to make it superfluous, but we will see that it can often be easier to test for (iii). The common, pre-theoretic understanding of expletives does not include idiom parts such as *the bucket* in *kick the bucket*, so it is necessary to restrict the concept further. Postal and Pullum (1988) do this by (i), which restricts expletives to be pro-forms. This is a relatively weak constraint on the form of expletives. We will see later that it may be desirable to strengthen this criterion and require expletives to be pro-forms that are selected by the predicate with which it occurs. Such purely formal selection is needed in many cases, since expletives are not interchangeable across constructions – for example, *there rains* is not an acceptable sentence of English. Criteria (ii) and (iii) from the definition of Postal and Pullum (1988) may be hard to apply directly in a UD setting, as UD is a syntactic, not a semantic, annotation framework. On the other hand, many decisions in UD are driven by the need to provide annotations that can serve as input for semantic analysis, and distinguishing between elements that do and do not refer and fill a thematic role therefore seems desirable.

In addition to the definition, Postal and Pullum (1988) provide tests for expletives. Some of these (tough-movement and nominalization) are not easy to apply cross-linguistically, but two of them are, namely absence of coordination and inability to license an emphatic reflexive.

(2) *It and John rained and carried an umbrella respectively.

(3) *It itself rained.

The inability to license an emphatic reflexive is probably due to the lack of referentiality. It is less immediately obvious what the absence of coordination diagnoses. One likely interpretation is that sentences like (2) are ungrammatical because the verb selects for a particular syntactic string as its subject. If that is so, form-selection can be considered a defining feature of expletives.

Finally, following Postal and Pullum (1988), we can draw a distinction between expletives that occur in chains and those that do not, where we understand a chain as a relation between an expletive and some other element of the sentence which has the thematic role that would normally be associated with the position of the expletive, for example, the subordinate clause in (4).

(4) It surprised me that she came.

It is not always possible to realize the other element in the chain in the position of the expletive. For example, the subordinate clause cannot be directly embedded under the preposition in (1).

Whether the expletive participates in a chain or not is relevant for the UD annotation insofar as it is often desirable – for the purposes of semantic interpretation – to give the semantically active element of the chain the "real" dependency label. For example, it is tempting to take the complement clause in (4) as the subject (`csubj` in UD) to stay closer to the semantics, although one is hard pressed to come up with independent syntactic evidence that an element in this position can actually be a subject. This is in line with many descriptive grammar traditions, where the expletive would be called the *formal* subject and the subordinate clause the *logical* subject.

We now review constructions that are regularly analyzed as involving an expletive in the theoretical literature and discuss these in the light of the definition and tests we have established.

2.1 Extraposition of Clausal Arguments

In many languages, verbs selecting a clausal subject or object often allow or require an expletive and place the clausal argument in extraposed position. In some cases, extraposition of the clausal argument is obligatory, as in (5) for English. Note that the clausal argument can be either a subject or an object, and thus the expletive in some cases appears in object position, as in (6). Also note that in so-called raising contexts, the expletive may actually be realized in the structural subject position of a verb governing the verb that selects the clausal argument (7).

(5) *It* seems that she came (en)

(6) Hij betreurt *het* dat jullie verliezen (nl)
He regrets it that you lose
'He regrets that you lose'

(7) *It* is going to be hard to sell the Dodge (en)

It is fairly straightforward to argue that this construction involves an expletive. Theoretically, *it* could be cataphoric to the following clause and so be referential, but in that case we would expect it to be able to license an emphatic reflexive. However, this is not what we find, as shown in (8-a), which contrasts with (8-b) where the raised subject is a referential pronoun.

(8) a. *It seems itself that she came
b. It seems itself to be a primary metaphysical principle

But if *it* does not refer cataphorically to the extraposed clause, its form must also be due to the construction in which it appears. This construction therefore fulfills the criteria of an expletive even on the strictest understanding.

2.2 Existential Sentences

Existential (or presentational) sentences are sentences that involve an intransitive verb and a noun phrase that is interpreted as the logical subject of the verb but does not occur in the canonical subject position, which is instead filled by an expletive. There is considerable variation between languages as to which verbs participate in this construction. For instance, while English is quite restrictive and uses this construction mainly with the copula *be*, other languages allow a wider range of verbs including verbs of position and movement, as illustrated in (9)–(11). There is also variation with respect to criteria for classifying the nominal constituent as a subject or object, with diagnostics such as agreement, case, and structural position often giving conflicting results. Some languages, like the Scandinavian languages, restrict the nominal element to indefinite nominals, whereas German for instance also allows for definite nominals in this construction.

(9) *Det* sitter en katt på mattan (sv)
it sits a cat on the-mat
'A cat sits on the mat'

(10) *Es* landet ein Flugzeug (de)
it lands a plane
'A plane lands'

(11) *Il* nageait quelques personnes (fr)
there swim some people
'Some people are swimming'

Despite the cross-linguistic variation, existential constructions like these are uncontroversial cases of expletive usage. The form of the pronoun(s) is fixed, it cannot refer to the other element of the chain for formal reasons, and no emphatic reflexive is possible.

2.3 Impersonal Constructions

By impersonal constructions we understand constructions where the verb takes a fixed, pronominal, argument in subject position that is not interpreted in semantics. Some of these involve zerovalent verbs, such as weather verbs, which are traditionally assumed to take an expletive subject in Germanic languages, as in Norwegian *regne* 'rain' (12). Others involve verb that also take a semantic argument, such as the French *falloir* in (13).

(12) *Det* regner (no)
it rains
'It is raining'

(13) *Il* faut trois nouveaux recrutements (fr)
it needs three new staff-members
'Three new staff members are needed'

Impersonal constructions can also arise when an intransitive verb is passivized (and the normal semantic subject argument therefore suppressed).

(14) *Es* wird gespielt (de)
It is played
'There is playing'

In all these examples, the pronouns are clearly non-referential, no emphatic reflexive is possible and the form is selected by the construction, so these elements can be classified as expletive.

2.4 Passive Reflexives

In some Romance and Slavic languages, a passive can be formed by adding a reflexive pronoun which does not get a thematic role but rather signals the passive voice.

(15) dospívá *se* dříve (cs)
mature REFL earlier
'(they/people) mature up earlier'

In Romance languages, as shown by Silveira (2016), these are not only used with a strictly passive meaning, but also with inchoative (anticausative) and medio-passive readings.

(16) La branche *s'* est cassée
The branch SE is broken
'The branch broke.'

In all of these cases, it is clear that the reflexive element does not receive a semantic role. In (15), *dospívá* 'mature' only takes one semantic argument, and in (16), the intended reading is clearly not that the branch broke itself. We conclude that these elements are expletives according to the definition above. This is in line with the proposal of Silveira (2016).

2.5 Inherent Reflexives

Many languages have verbs that obligatorily select a reflexive pronoun without assigning a semantic role to it:

(17) Pedro *se* confundiu (pt)
Pedro REFL confused
'Pedro was confused'

(18) Směje *se* (cs)
laugh REFL
'he/she/it laughs'

There are borderline cases where the verb in question can also take a regular object, but the semantics is subtly different. A typical case are verbs like *wash*. That there are in fact two different interpretations is revealed in Scandinavian by the impossibility of coordination. (19) is grammatical unless *seg* is stressed.

(19) *Han vasket seg og de andre (no)
He washed REFL and the others
'He washed himself and the others'

From the point of view of our definition, it is clear that inherent reflexives (by definition) do not receive a semantic role. It may be less clear that they are non-referential: after all, they typically agree with the subject and could be taken to be co-referent. It is hard to test for non-referentiality in the absence of any semantic role. In particular, the emphatic reflexive test is not easily applicable, since it may be the subject that antecedes the emphatic reflexive in cases like (20).

(20) Elle s'est souvenue elle-même
she REFL-is reminded herself
'She herself remembered...'

Inherent reflexives agree with the subject, and thus their form is not determined (only) by the verb. Nevertheless, under the looser understanding of the formal criterion, it is enough that reflexives are pronominal and thus can be expletives. This is also the conclusion of Silveira (2016).

2.6 Clitic Doubling

The UD guidelines explicitly mention that "true" (that is, regularly available) clitic doubling, as in the Greek example in (21), should be annotated using the `expl` relation:

(21) pisteuô oti einai dikaio na to
I-believe that it-is fair that this-CLITIC
anagnôrisoume auto (el)
we-recognize this

The clitic *to* merely signals the presence of the full pronoun object and it can be argued that it is the latter that receives the thematic role. It is less clear, however, that *to* is non-referential, hence it is unclear that this is an instance of an expletive. The alternative is to annotate the clitic as a core argument and use `dislocated` for the full pronoun (as is done for other cases of doubling in UD).

3 Expletives in UD 2.1 treebanks

We will now present a survey of the usage of the `expl` relation in current UD treebanks. In particular, we will relate the constructions discussed in Section 2 to the treebank data. Table 1 gives an overview of the usage of `expl` and its language specific extensions in the treebanks in UD v2.1.[1] We find that, out of the 60 languages included in this release, 27 make use of the `expl` relation, and its use appears to be restricted to European languages. For those languages that have multiple treebanks, `expl` is not always used in all treebanks (Finnish, Galician, Latin, Portuguese, Russian, Spanish). The frequency of `expl` varies greatly, ranging from less than 1 per 1,000 words (Catalan, Greek, Latin, Russian, Spanish, Ukranian) to more than 2 per 100 words (Bulgarian, Polish, Slovak). For most of the languages, there is a fairly limited set of lemmas that realize the `expl` relation. Treebanks with higher numbers of lemmas are those that label inherent reflexives as `expl` and/or do not always lemmatize systematically. Some treebanks not only use `expl`, but also the subtypes `expl:pv` (for inherent reflexives), `expl:pass` (for certain passive constructions), and `expl:impers` (for impersonal constructions).

The counts and proportions for specific constructions in Table 1 were computed as follows. Extraposition covers cases where an expletive cooccurs with a `csubj` or `ccomp` argument as in the top row of Figure 1. This construction occurs frequently in the Germanic treebanks (Dutch, English, German, Norwegian, Swedish), as in (22), but is also fairly frequent in French treebanks, as in (23).

(22) *It* is true that Google has been in acquisition mode (en)

(23) Il est de notre devoir de participer [. . .] (fr)
it is of our duty to participate [. . .]
'It is our duty to participate . . . '

Existential constructions can be identified by the presence of a nominal subject (`nsubj`) as a sibling of the `expl` element, as illustrated in the middle row of Figure 1. Existential constructions are very widespread and span several language families in the treebank data. They are common in all Germanic treebanks, as illustrated in (24), but are also found in Finnish, exemplified in (25), where these constructions account for half of all expletive occurrences, as well as in several Romance languages (French, Galician, Italian, Portuguese), some Slavic languages (Russian and Ukrainian), and Greek.

(24) *Es* fehlt ein System umfassender sozialer
it lacks a system comprehensive social
Sicherung (de)
security
'A system of comprehensive social security is lacking'

(25) *Se* oli paska homma, että Jyrki loppu (fi)
it was shit thing that Jyrki end
'It was a shit thing for Jyrki to end'

For the impersonal constructions discussed in Section 2.3, only a few UD treebanks make use of an explicit `impers` subtype (Italian, Romanian). Apart from these, impersonal verbs like *rain* and French *falloir* prove difficult to identify reliably across languages using morphosyntactic criteria. For impersonal passives, on the other hand, there are morphosyntactic properties that we may employ in our survey. Passives in UD are marked either morphologically on the verb (by the feature `Voice=Passive`) or by a passive auxiliary dependent (`aux:pass`) in the case of periphrastic passive constructions. These two passive constructions are illustrated in the bottom row (left

[1] The raw counts as well as the script we used to collect the data can be found at `github.com/gossebouma/expletives`

	Banks	Count	Freq	Lemmas	Extraposed		Existential		Impersonal		Reflexives		Remaining	
Bulgarian		3379	0.021	7	12	0.0	82	0.02	2	0.0	3204	0.95	79	0.02
Catalan		512	0.001	4	0	0.0	0	0.0	0	0.0	512	1.0	0	0.0
Croatian		2173	0.011	11	2	0.0	4	0.0	1	0.0	2161	0.99	5	0.0
Czech	5/5	35929	0.018	4	0	0.0	0	0.0	0	0.0	35929	1.0	0	0.0
Danish		441	0.004	2	8	0.02	10	0.02	62	0.14	0	0.0	361	0.82
Dutch	2/2	459	0.001	5	321	0.7	120	0.26	6	0.01	0	0.0	12	0.03
English	4/4	1221	0.003	6	380	0.31	724	0.59	9	0.01	0	0.0	107	0.09
Finnish	1/3	524	0.003	9	15	0.03	268	0.51	53	0.1	0	0.0	188	0.36
French	5/5	6117	0.005	26	162	0.03	1486	0.24	27	0.0	3378	0.55	1064	0.17
Galician	1/2	288	0.01	6	19	0.07	131	0.45	0	0.0	0	0.0	138	0.48
German	2/2	487	0.003	1	114	0.23	287	0.59	21	0.04	1	0.0	64	0.13
Greek		18	0.000	1	0	0.0	6	0.33	0	0.0	0	0.0	12	0.67
Italian	4/4	4214	0.009	22	107	0.03	1901	0.45	589	0.14	396	0.09	1218	0.29
Latin	1/3	257	0.001	1	0	0.0	0	0.0	0	0.0	257	1.0	0	0.0
Norwegian	3/3	6890	0.01	8	1894	0.27	1758	0.26	374	0.05	0	0.0	2864	0.42
Polish		1708	0.02	1	0	0.0	0	0.0	6	0.0	1702	1.0	0	0.0
Portuguese	2/3	1624	0.003	1	20	0.01	628	0.39	20	0.01	672	0.41	284	0.17
Romanian	2/2	5209	0.002	22	43	0.01	327	0.06	140	0.03	4281	0.82	418	0.08
Russian	2/3	55	0.000	3	6	0.11	42	0.76	1	0.02	0	0.0	6	0.11
Slovak		2841	0.03	3	0	0.0	0	0.0	0	0.0	2841	1.0	0	0.0
Slovenian	2/2	2754	0.02	2	0	0.0	1	0.0	0	0.0	2297	1.0	0	0.0
Spanish	1/3	503	0.001	2	0	0.0	0	0.0	0	0.0	503	1.0	0	0.0
Swedish	3/3	1079	0.005	6	371	0.34	283	0.26	85	0.08	0	0.0	340	0.32
Ukrainian		94	0.001	4	16	0.17	62	0.66	0	0.0	12	0.13	4	0.04
Upper Sorbian		177	0.02	1	0	0.0	0	0.0	1	0.01	176	0.99	0	0.0

Table 1: Use of `expl` in UD v2.1 treebanks. Languages with Count $<$ 10 left out (Arabic, Sanskrit). Freq = average frequency for treebanks containing `expl`. Count and proportion for construction types.

and center) of Figure 1. The quantitative overview in Table 1 shows that impersonal constructions occur mostly in Germanic languages, such as Danish, German, Norwegian and Swedish, illustrated by (26). These are all impersonal passives. We note that both Italian and Romanian also show a high proportion of impersonal verbs, due to the use of `expl:impers` mentioned above and exemplified by (27).

(26) *Det* ble ikke nevnt hvor omstridt
it was not mentioned how controversial
han er (no)
he is
'It was not mentioned how controversial he is'

(27) *Si* compredono inoltre i figli adottivi
it includes also the children adopted
(it)
'Adopted children are also included'

Both the constructions of passive reflexives and inherent reflexives (Sections 2.4 and 2.5), make use of a reflexive pronoun. Some treebanks distinguish these through subtyping of the `expl` relation, for instance, `expl:pass` and `expl:pv` in the Czech treebanks. This is not, however, the case across languages and since the reflexive passive does not require passive marking on the verb, it is difficult to distinguish these automatically based on morphosyntactic criteria. In Table 1 we therefore collapse these two construction types (Reflexive). In addition to the `pv` subtype, we further rely on another morphological feature in the treebanks in order to identify inherent reflexives, namely the `Reflex` feature, as illustrated by the Portuguese example in Figure 1 (bottom right).[2] In Table 1 we observe that the distribution of passive and inherent reflexives clearly separates the different treebanks. They are highly frequent in Slavic languages (Bulgarian, Croatian, Czech, Polish, Slovak, Slovenian, Ukrainian and Upper Sorbian). as illustrated by the passive reflexive in (28) and the inherent reflexive in (29). They are also frequent in two of the French treebanks and in Brazilian Portuguese. Interestingly, they are also found in Latin, but only in the treebank based on medieval texts.

(28) O centrální výrobě tepla *se* říká,
about central production heating it says
že je nejefektivnější (cs)
that the most-efficient

[2] The final category discussed in section 2 is that of clitic doubling. It is not clear, however, how one could recognize these based on their morphosyntactic analysis in the various treebanks and we therefore exclude them from our empirical study, although a manual analysis confirmed that they exist at least in Bulgarian and Greek.

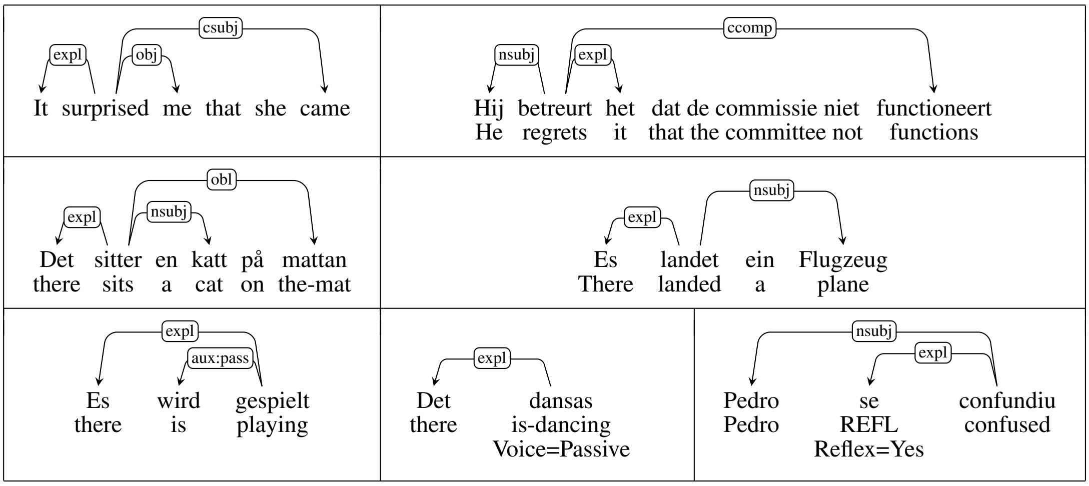

Figure 1: UD analyses of *extraposition* [(4) and (6)] (top), *existentials* [(9) and (10)] (middle), *impersonal constructions* (bottom left and center), and *inherent reflexives* [(17)] (bottom right).

‘Central heat production is said to be the most efficient’

(29) Skozi steno slišim, kako *se*
through wall I-hear-it, how REFL
zabavajo. (sl)
have-fun
‘I hear through the wall how they have fun’

(30) O deputado *se* aproximou (pt)
the deputy REFL approached
‘The deputy approached’

It is clear from the discussion above that all constructions discussed in Section 2 are attested in UD treebanks. Some languages have a substantial number of expl occurrences that are not captured by our heuristics (i.e. the Remaining category in Table 1). In some cases (i.e. Swedish and Norwegian), this is due to an analysis of *cleft* constructions where the pronoun is tagged as expl. It should be noted that the analysis of clefts differs considerably across languages and treebanks, and therefore we did not include it in the empirical overview. Another frequent pattern not captured by our heuristics involves clitics and clitic doubling. This is true especially for the Romance languages, where Italian and Galician have a substantial number of occurrences of expl marked as Clitic not covered by our heuristics. In French, a frequent pattern not captured by our heuristics is the *il y a* construction.

The empirical investigation also makes clear that the analysis of expletives under the current UD scheme suffers from inconsistencies. For inherent reflexives, the treebanks for Croatian, Czech, Polish, Portuguese, Romanian, and Slovak use the subtype expl:pv, while the treebanks for French, Italian and Spanish simply use expl for this purpose. And even though languages like German, Dutch and Swedish do have inherent reflexives, their reflexive arguments are currently annotated as regular objects.

Even in different treebanks for one and the same language, different decisions have sometimes been made, as is clear from the column labeled Banks in Table 1. Of the three treebanks for Spanish, for instance, only Spanish-AnCora uses the expl relation, and of the three Finnish UD treebanks, only Finnish-FTB. In the French treebanks, we observe that the expl relation is employed to capture quite different constructions. For instance, in French-ParTUT, it is used for impersonal subjects (non-referential *il*, whereas the other French treebanks do not employ an expletive analysis for these. We also find that annotation within a single treebank is not always consistent. For instance, whereas the German treebank generally marks *es* in existential constructions with *geben* as expl, the treebank also contains a fair amount of examples with *geben* where es is marked nsubj, despite being clearly expletive.

4 Towards Consistent Annotation of Expletives in UD

Our investigations in the previous section clearly demonstrate that expletives are currently not annotated consistently in UD treebanks. This is partly due to the existence of different descriptive and theoretical traditions and to the fact that many treebanks have been converted from annotation schemes that differ in their treatment of expletives. But the situation has probably been made worse by the lack of detailed guidelines concerning which constructions should be analyzed as involving expletives and how exactly these constructions should be annotated. In this section, we will take a first step towards improving the situation by making specific recommendations on both of these aspects.

Based on the definition and tests taken from Postal and Pullum (1988), we propose that the class of expletives should include non-referential pro-forms involved in the following types of constructions:

1. Extraposition of clausal arguments (Section 2.1)
2. Existential (or presentational) sentences (Section 2.2)
3. Impersonal constructions (including weather verbs and impersonal passives) (Section 2.3)
4. Passive reflexives (Section 2.4)
5. Inherent reflexives (Section 2.5)

For inherent reflexives, the evidence is not quite as clear-cut as for the other categories, but given that the current UD guidelines recommend using `expl` and given that many treebanks already follow these guidelines, it seems most practical to continue to include them in the class of expletives, as recommended by Silveira (2016). By contrast, the arguments for treating clitics in clitic doubling (Section 2.6) as expletives appears weaker, and very few treebanks have implemented this analysis, so we think it may be worth reconsidering their analysis and possibly use `dislocated` for all cases of double realization of core arguments.

The distinction between core arguments and other dependents of a predicate is a cornerstone of the UD approach to syntactic annotation. Expletives challenge this distinction by (mostly) behaving as core arguments syntactically but not semantically. In chain constructions like extraposition and existentials, they compete with the other chain element for the core argument relation. In impersonal constructions and inherent reflexives, they are the sole candidate for that relation. This suggests three possible ways of treating expletives in relation to core arguments:

1. Treat expletives as distinct from core arguments and assign the core argument relation to the other chain element (if present).
2. Treat expletives as core arguments and allow the other chain element (if present) to instantiate the same relation (possibly using subtypes to distinguish the two).
3. Treat expletives as core arguments and forbid the other chain element (if present) to instantiate the same relation.

All three approaches have advantages and drawbacks, but the current UD guidelines clearly favor the first approach, essentially restricting the application of core argument relations to *referential* core arguments. Since this approach is already implemented in a large number of treebanks, albeit to different degrees and with considerable variation, it seems practically preferable to maintain and refine this approach, rather than switching to a radically different scheme. However, in order to make the annotation more informative, we recommend using the following subtypes of the `expl` relation:

1. `expl:chain` for expletives that occur in chain constructions like extraposition of clausal arguments and existential or presentational sentences (Section 2.1–2.2)
2. `expl:impers` for expletive subjects in impersonal constructions, including impersonal verbs and passivized intransitive verbs (Section 2.3)
3. `expl:pass` for reflexive pronouns used to form passives (Section 2.4)
4. `expl:pv` for inherent reflexives, that is, pronouns selected by pronominal verbs (Section 2.5)

The three latter subtypes are already included in the UD guidelines,although it is clear that they are not used in all treebanks that use the `expl` relation. The first subtype, `expl:chain`, is a novel proposal, which would allow us to distinguish constructions where the expletive is dependent on the presence of a referential argument. This subtype could possibly be used also in clitic doubling, if we decide to include these among expletives.

5 Conclusion

Creating consistently annotated treebanks for many languages is potentially of tremendous importance for both NLP and linguistics. While our study of the annotation of expletives in UD shows that this goal has not quite been reached yet, the development of UD has at least made it possible to start investigating these issues on a large scale. Based on a theoretical analysis of expletives and an empirical survey of current UD treebanks, we have proposed a refinement of the annotation guidelines that is well grounded in both theory and data and that will hopefully lead to more consistency. By systematically studying different linguistic phenomena in this way, we can gradually approach the goal of global consistency.

Acknowledgments

We are grateful to two anonymous reviewers for constructive comments on the first version of the paper. Most of the work described in this article was conducted during the authors' stays at the Center for Advanced Study at the Norwegian Academy of Science and Letters.

References

Hans Bennis. 2010. *Gaps and dummies*. Amsterdam University Press.

Adriane Boyd, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential it: A machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, FeatureEng '05, pages 40–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anna Cardinaletti. 1997. Agreement and control in expletive constructions. *Linguistic Inquiry*, 28(3):521–533.

Richard Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and linguistic computing*, 16(1):45–58.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. Parcor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *9th International Conference on Language Resources and Evaluation (LREC), May 26-31, 2014, Reykjavik, Iceland*, pages 3191–3198. European Language Resources Association.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16.

Ilan Hazout. 2004. The syntax of existential constructions. *Linguistic Inquiry*, 35(3):393–430.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? disambiguating the different readings of the pronoun 'it'. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16. Springer.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *10th International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia*, pages 1659–1666. European Language Resources Association.

Christer Platzack. 1987. The Scandinavian languages and the null-subject parameter. *Natural Language & Linguistic Theory*, 5(3):377–401.

Paul M Postal and Geoffrey K Pullum. 1988. Expletive noun phrases in subcategorized positions. *Linguistic Inquiry*, 19(4):635–670.

Natalia Silveira. 2016. *Designing Syntactic Representations for NLP: An Empirical Investigation*. Ph.D. thesis, Stanford University, Stanford, CA.

Olga Uryupina, Mijail Kabadjov, and Massimo Poesio. 2016. Detecting non-reference and non-anaphoricity. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources, and Applications*, pages 369–392. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sten Vikner. 1995. *Verb movement and expletive subjects in the Germanic languages*. Oxford University Press on Demand.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using coreference links to improve Spanish-to-English machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40.