Core Arguments in Universal Dependencies

Daniel Zeman
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia
zeman@ufal.mff.cuni.cz

Abstract

We investigate how core arguments are coded in case-marking Indo-European languages. Core arguments are a central concept in Universal Dependencies, yet it is sometimes difficult to match against terminologies traditionally used for individual languages. We review the methodology described in (Andrews, 2007), and include brief definitions of some basic terms. Statistics from 26 UD treebanks show that not all treebank providers define the core-oblique boundary the same way. Therefore we propose some refinement and particularization of the guidelines that would improve cross-treebank consistency on the one hand, and be more sensitive to the traditional grammar on the other.

1 Introduction

The opposition of core vs. oblique dependents is one of the central concepts in Universal Dependencies (Nivre et al., 2016); this distinction is intentionally preferred to the argument/adjunct distinction. However, difficulties in recognizing core arguments in individual languages, combined with often incompatible traditional terminology, have led to confusion and data inconsistency. UD documentation has greatly improved since its version 1 and provides now a list of potential criteria that may help to draw the core vs. oblique borderline; however, it is still just a set of hints, not a definition. The English UD uses a relatively simple rule: as soon as a preposition is involved, the noun phrase cannot be analyzed as a core argument. Unfortunately, there are many languages where the situation is more complex. In the present work we are particularly interested in languages that use both case morphology and prepositions to mark arguments.

We review one possible universal methodology to identify coding of core arguments, and show how it applies to these languages. Terms like argument, transitive verb or indirect object are often taken for known and granted (both in the UD guidelines and in the literature) but the problem is that their definition may differ by language or by author, and it is not easy to see how they work across languages. Therefore we briefly define the necessary terms as well.

2 Core Arguments in Language Typology

In this section we provide a brief definition of core arguments; for a much more detailed discussion see (Andrews, 2007), which is our primary source.

2.1 Arguments and Adjuncts

Arguments are noun phrases that fulfill semantic roles determined by verbs, or more generally by predicates. Depending on language, the verb may also specify requirements on the position of the individual arguments and on their form, such as morphological case marking or preposition.

In contrast, adjuncts are noun phrases that specify additional circumstances such as location, time and manner. Neither their form nor their meaning is determined by the verb. They can accompany any predicate; some collocations may be difficult to interpret semantically but they are not ungrammatical. Likewise, the form of the adjuncts is determined by their meaning rather than by the verb.

Hence, the phrase marked by the preposition *on* is an argument in *I rely on him* or in *I will act on the matter*, but it is an adjunct in *I will work on Saturday* or *I live on an island*. These examples are relatively easy to understand; however, in general the argument-adjunct distinction is not always trivial, and UD avoids it (from the guidelines: "We take the distinction to be sufficiently subtle (and its existence as a categorical distinction sufficiently questionable) that the best practical solution is to

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 287-296,
Pisa, Italy, September 18-20 2017

eliminate it.") Nevertheless, we will see that even for the distinction between core and oblique arguments, it is sometimes necessary to make sure that the noun phrase is actually an argument and not an adjunct. Whenever we say 'argument' in the rest of the paper, we think of it as defined in the beginning of this section.

2.2 Transitive Verbs

The most reliable means of distinguishing between core and oblique arguments are the encoding strategies such as word order, adpositions and morphological case. However, the strategies are always specific to a language and cannot be used in a cross-linguistically applicable definition. Therefore we start with semantic roles to identify *prototypical core arguments*, then we observe the strategies that the language uses to mark them, and finally generalize to other arguments using the same strategy, despite their semantic roles being different from the prototypical core arguments.

The prototypical core can be observed with *primary transitive verbs,* i.e. verbs that take two arguments whose semantic roles are *agent* and *patient,* respectively. The agent, typically an animate entity, is responsible for an action, and the patient is directly affected by the action. *To kill* is an example of a primary transitive verb: in *George killed the dragon,* George is the agent who did the killing (note that it is not necessary for an agent to act willingly; it could also be an accident). Without any doubt, the dragon is the entity most affected by the killing, and the killing caused a change of the dragon's state. Hence the dragon qualifies as the patient.

Languages differ in how they make clear who killed whom. In English, it is the position of the arguments relative to the verb. In Czech, the agent would be in its nominative form, and the patient in the accusative.[1] However, in good many languages the same coding strategy is also used with verbs whose two arguments have other semantic roles. For instance, *to love* takes two arguments but it is not a *primary* transitive verb because the roles of the arguments are better described as "experiencer" and "goal" rather than "agent" and "patient". Nevertheless, the verb is transitive in both English and Czech because the two arguments are marked in these languages in exactly the same way as the arguments of *to kill*.

[1]Unless the verb is in its passive form.

Following (Andrews, 2007), if a noun phrase is serving as an argument of a two-argument verb, and receiving a morphological and syntactic treatment normally accorded to an agent of a primary transitive verb, it has the **grammatical function A;** analogically, an argument receiving treatment normally accorded to a patient of a primary transitive verb has the grammatical function **P.**[2]

2.3 Intransitive Verbs

If a verb takes just a single argument, the verb is called *intransitive* and its argument has the grammatical function **S.** Depending on language (and in some languages depending on individual verbs), the S argument of intransitive verbs may conform to the same grammatical rules as the A argument of transitive verbs, or as the P argument, or it can be different from both A and P.

2.4 Core and Oblique Arguments in UD

S, A and P are considered core grammatical functions (Andrews, 2007, p. 164). As UD refers to Andrews,[3] we can project to UD: Arguments that have one of the **S/A/P functions are core arguments.** Nominals whose grammatical function is A or S are called *subjects* and their dependency relation to the verb is `nsubj`. Nominals whose grammatical function is P are called (direct) *objects* and their dependency relation to the verb is `obj`. Both subject and object are considered *core arguments*. In addition, UD uses a special relation `iobj` for what it calls *indirect objects;* we will investigate them in Section 4.

Using the concepts defined so far, it is now possible to lay down rules for core arguments in individual languages. For instance, in English, if a bare noun phrase (i.e. without a preposition) is an argument of a verb, it is a core argument; if it occurs in a simple declarative clause and precedes the verb, it is its subject; if it follows the verb, it is an object. Note the important condition *if it is an argument, not adjunct*. While adjuncts usually take prepositions in English, they occasionally appear as bare noun phrases too; as an example, consider

[2]Note that some authors use the terms *agent* and *patient* to refer to what we call A and P here, rather than to the semantic roles; cf. the functors on the t-layer of the Prague Dependency Treebank (Hajič et al., 2006). It is important not to confuse that terminology with ours: for example, the two arguments of *to love* would then be agent and patient, while we argue that they are not.

[3]`http://universaldependencies.org/u/overview/syntax.html`, retreived 2017-07-23

the temporal adjunct *this week* in *I am not working this week.*

On the other hand, verbs in many languages have arguments that are marked by coding strategies that are also used by adjuncts, but that are different from strategies used by core arguments. Such arguments are called *oblique*. For instance, the second argument of *act* in *I will act on the matter* is marked by the preposition *on*. Since core arguments in English do not take prepositions, this is an oblique argument. In UD, both oblique arguments and adjuncts are attached to the verb via an `obl` relation (if they are noun phrases).

Note that the methodology described in this section is not the only possible. (Dixon, 2012, vol. 1 sec. 3.2 and vol. 2 sec. 13) defines core arguments as those that "must be either stated or understood from the context;" the opposite of core are *peripheral arguments*. Dixon's core arguments are in spirit similar to those of Andrews, but his definition does not guarantee that no verbs have their core arguments marked by "oblique" strategies.

3 Languages with Case-Marking Morphology

A number of Indo-European languages have the morphological category of case. In these languages, the most typical coding of core arguments is the nominative case (subject) and the accusative case (object). However, there are usually more cases than these two, and the question arises whether arguments in other morphological cases count as core arguments. (Andrews, 2007) gives an example from German: the verb *helfen* ("to help") takes two arguments, one in nominative and the other in dative. We can say that *helfen* is a primary transitive verb because the roles of the two arguments are agent and patient. It can also be passivized, which is a typical property of transitive verbs; however, unlike verbs with accusative objects, the dative argument of *helfen* stays in the dative and does not become subject when the verb appears in the passive voice. We thus have an argument whose grammatical behavior is not identical with the more typical accusative object, yet it is sufficiently similar to qualify as a core argument. In consequence, all arguments that are bare nominals in dative are core arguments in German.[4]

A similar observation can be made in Slavic languages. In fact everything that we just said about the German verb *helfen* also applies to the Czech verb *pomoci* ("to help"). However, Czech has more cases than German, and there are two-argument verbs whose second argument is neither accusative nor dative. Bare genitives and instrumentals may act as arguments too; moreover, there are prepositional arguments in genitive, dative, accusative, locative or instrumental. Many of these verbs can be passivized in the same way as *pomoci*. For example, the verb *hýbat* ("to move") takes an instrumental patient-object: in *Martin hýbá nábytkem* "Martin moves the furniture", the noun *nábytek* ("furniture") takes its instrumental form. When passivized, the agent disappears and the patient stays in instrumental: *Nábytkem bylo hýbáno* "The furniture has been moved."

A somewhat different example is the verb *dotknout se* ("to touch"). This verb is inherently reflexive, i.e. obligatorily accompanied by the reflexive pronoun *se*.[5] It takes two arguments: the agent is in nominative as usual, and the patient is in genitive. According to the semantic roles we could argue that it is a primary transitive verb. However, reflexive verbs cannot be passivized in Czech: **Bylo se ho dotknuto* ("He has been touched") is not grammatical. Thus we have a two-argument verb whose arguments pass the tests on coreness laid out in Section 2, yet it does not permit passivization, an operation usually associated with transitive verbs (note however that passivization is not universal and cannot be added as a requirement for transitive verbs).

So we have three types of transitive verbs w.r.t. passivization (1. accusative; 2. non-accusative non-reflexive; 3. reflexive). We can also observe varying degree of coreness. The largest proportion of *primary* transitive verbs will indisputably be found among verbs with accusative objects. Verbs taking objects in genitive, dative and instrumental often select roles quite different from the (proto-) patient; only a handful can be regarded as primary transitive verbs. Even harder to find are patients among prepositional arguments, but some of them would deserve to be at least considered as candi-

[4]Note that this finding is not without controversy. Some authors classify the German dative as an oblique case, although they do not specify what are the properties their classification is based on (Foley, 2007, p. 377).

[5]With inherently reflexive verbs, the reflexive pronoun (sometimes termed particle), although syntactically autonomous, is part of the verbal lexeme, not an argument. However, transitive verbs can take reflexive pronouns as their objects.

dates. At the same time, bare accusative is very rarely used for adjuncts, which are slightly more common among other bare noun phrases, and the majority of them are prepositional phrases.

Strictly following the tests from Section 2 and from (Andrews, 2007), all Czech arguments would be core and none of them would be oblique. While this "classification" aligns with the notion of objects in the Czech grammar (see Section 5), it is of no benefit. It does not make sense to delimit the core of a set if it comprises the entire set; furthermore, the identification of core arguments would now be reduced to the argument-adjunct distinction, which UD wanted to eliminate.

So, is there a way to interpret Section 2 with less extreme results? There is one word that may provide the remedy. In 2.2 we say that the P function is recognized by treatment *normally* accorded to a patient of a primary transitive verb. Now we showed that bare accusative is the "most normal" coding strategy and prepositional phrases are still possible, but arguably "least normal" for patients. Out of the three possible coding strategies (bare accusatives, bare non-accusatives and prepositional phrases), we could decide that one or two are not normal enough. Our cross-linguistic detection of core arguments will become a bit less deterministic but more flexible; it may be the right compromise to use.

4 Ditransitive Verbs and Indirect Objects

Predicates may define more than two roles. In the Czech sentence *Firma mu zvýšila plat z dvaceti na třicet tisíc korun* lit. "Company him raised salary from twenty to thirty thousand crowns" (Lopatková et al., 2016) the verb *zvýšit* ("raise") has four or five arguments.[6] With an extreme interpretation of Section 2 we could even claim that all of them are core arguments. It is usually not assumed that languages have that many core arguments; nevertheless, it is accepted that some verbs in some languages have three. Such verbs are called *ditransitive*.

Verbs of giving, taking and related concepts (e.g. teaching = giving knowledge) are prototypical examples in many languages. Their arguments correspond to the semantic roles of agent, theme (or patient) and recipient (Dryer, 2007). In terms of grammatical relations they correspond to subject, direct and indirect object. There is a confusion potential though. Some grammars will define indirect object as the argument with the recipient role. However, this argument is not necessarily a core argument by our definition: in English in *John gave Mary a flower*, the recipient *(Mary)* is a core argument; but in *John gave a flower to Mary*, the recipient is oblique. When we restrict ourselves to core arguments, there are clearly languages and verbs with two objects but it is less clear whether (and why) one of them deserves a special term. (Andrews, 2007) notes that "the status of the notion of 'indirect object' is problematic and difficult to sort out. The top priority is to work out what properties recipients and themes do and do not share with P arguments of primary transitive verbs."

In Universal Dependencies, the v2 guidelines say that "The indirect object of a verb is any nominal phrase that is a core argument of the verb but is not its subject or (direct) object." Such a definition is not sufficient for us—any core argument that is not a subject is an object. The UD guidelines "define" the (direct) object as the second most core argument after subject. They do not provide means to quantify coreness, though. For our group of languages, we could use the observation from Section 3 that there are three coding strategies ordered by decreasing convincingness of their core status. However, UD also assumes that the relation `iobj` is only used with predicates that have more than one object, i.e., the indirect object cannot exist without a direct one. This rule would have to be changed, otherwise we cannot say that all bare dative arguments are `iobj`. For example, the German verb *helfen* does not have any accusative object that could be labeled `obj`.

5 Traditional Terminology

Traditional grammars in good many languages use less restrictive definitions of object than UD. It is not unusual to encounter non-accusative and even prepositional objects, no matter of their status as core or oblique arguments.

The school grammar of Czech (Havránek and Jedlička, 1966) is a concise but respected piece of work, which does not diverge from the mainstream terminology used by linguists. It provides a definition of object that is identical to our definition of argument in Section 2.1. Indirect object is mentioned only briefly as a possible name

[6] Depending on whether the beneficiary *him* is accepted as argument rather than adjunct.

	Nom	Acc	Dat	Gen	Abl	Loc	Ins	Voc	None
be	**36**/0	**20**/8	**2**/1	**7**/8		0/12	**3**/3		
bg	**9**/**1**	**14**/**1**	**3**/0						**46**/**27**
cs	**29**/0	**29**/**5**	**5**/**2**	**3**/**7**		0/**12**	**3**/**3**		**2**/**1**
cs_2	**27**/0	**31**/**6**	**4**/**2**	**3**/**7**		0/**12**	**4**/**3**		**1**/0
cs_3	**32**/0	**27**/**4**	**1**/**3**	**2**/10		0/**12**	**5**/**2**		1/0
cu	**26**/0	**21**/9	**15**/4	**7**/4		0/7	**2**/2	2/0	
de	**35**/0	**19**/6	**3**/20	0/1					**6**/11
el	**35**/0	**29**/29		**1**/2				1/0	**2**/1
got	**28**/0	**26**/6	**15**/20	**2**/1				1/0	
grc	**26**/0	**34**/7	**14**/5	**6**/6				2/0	
grc_2	**23**/0	**31**/11	**14**/6	**5**/8				1/0	**1**/0
hr	**32**/0	**30**/7	**3**/0	**4**/7		0/10	**2**/2		
la	**24**/0	**33**/8	**8**/0		**16**/9			1/0	
la_2	**36**/0	**19**/**15**	**5**/0		**5**/**19**	1/0			0/1
la_3	**24**/0	**31**/11	**9**/0	**1**/0	**9**/13			1/0	**1**/0
lt	**33**/0	**22**/6	**5**/0	**11**/5		7/0	**7**/1		**2**/1
lv	**37**/0	**21**/**6**	**8**/**5**	**2**/**4**		15/0			**1**/0
pl	**29**/0	**20**/**7**	**4**/0	**5**/**8**		0/**10**	**3**/**3**		**10**/0
pt	**1**/0	**6**/0	**1**/0						**57**/35
ru	**29**/0	**15**/8	**3**/4	**5**/8		0/19	**6**/3		
ru_2	**34**/0	**20**/7	3/3	**5**/7		0/11	6/3		
sa	**43**/0	**30**/0	1/0	4/0	**3**/0	6/0	**9**/0	**3**/0	**1**/0
sk	**27**/0	**24**/**6**	**6**/**2**	**1**/6		0/9	**2**/**3**		**14**/0
sl	**22**/0	**24**/8	**6**/1	**6**/4		0/14	0/6		**10**/0
sl_2	**25**/1	**25**/7	**6**/0	**6**/3		0/10	0/4		**14**/0
uk	**33**/0	**22**/9	**4**/0	**4**/10		0/10	4/3		

Table 1: Distribution (percentage) of morphological cases found at nominal dependents of verbs. Both occurrences with / without adposition are counted. Only Indo-European languages with three or more cases in UD 2.0 are shown. Languages are identified by their ISO 639 codes; when there are multiple treebanks per language, numerical indices are used instead of identifiers for brevity. **Highlight red** = mostly core relations (including `expl`). **Highlight blue** = mostly oblique, but significant (10% or more) amount of core also present.

for the dative argument of ditransitives. Textbooks use a question test to distinguish objects from non-objects. If a dependent of the verb can be queried by an interrogative adverb *(where, when, how)*, or by one of a few additional expressions such as *for what purpose*, it is an adverbial modifier—even if realized as a noun phrase! If we must use an interrogative pronoun *(who, what)* it is either a subject (if the pronoun is in nominative) or an object (otherwise). Thus in *spoléhám na kamarády* ("I rely on friends"), the prepositional phrase is object because the only plausible question is with a pronoun: *Na koho spoléhám?* ("Who do I rely on?"). In contrast, the prepositional phrase in *pojedu na Slovensko* ("I will go to Slovakia") is not normally queried by **Na co pojedu?* "What will I go to?" Instead, we use an adverb and ask *Kam pojedu?* "Where will I go?" Thus this phrase is not an object. If objects are defined this way, then most objects are arguments and most adverbials are adjuncts; the notion of core arguments does not play a role.

According to (Karlík et al., 2016), some more detailed grammar descriptions do distinguish indirect objects but they still do not restrict objects to core arguments. Bare accusative objects are direct (even in the rare cases when a verb has two accusative objects). Objects in other cases, including prepositional objects, are indirect (even with verbs like *pomoci* "to help" where no direct object is possible). A verb is transitive if it takes a direct object. Looking back at Section 3, we see that these direct

objects are always core arguments and they belong to the most core-like subset. Indirect objects may or may not be core arguments depending on how strictly we follow the principles from Section 2.

Such a perspective is not specific to Czech; it is rather dominant in European linguistics.

In their comparative grammar of Slavic languages, (Sussex and Cubberley, 2006, p. 339, 351–352) use the term *transitive verb* for verbs whose object is a bare noun phrase in any case; verbs with prepositional objects are neither transitive nor intransitive. *Direct object* is a synonym for bare accusative; other objects are referred to as *non-accusative objects* and *prepositional objects*. *Indirect object* seems to be used just for the semantic role of recipient (expressed by bare dative), probably assuming that the English readership will find the term familiar.

Another example, this time outside the Slavic group, is the cannonical grammar of German. (Helbig and Buscha, 1998, p. 53 and 545) distinguish accusative object, dative object, genitive object and prepositional object. Adjunct-like noun phrases are considered adverbial modifiers. Transitive verbs are those that take an accusative object and this object can become subject in a passive clause. Verbs that take an accusative object but cannot be passivized (*enthalten* "contain," *bekommen* "get" etc.) are called *medial verbs (Mittelverben)*. Intransitive verbs are those that do not take an accusative object, regardless whether they take a non-accusative object, prepositional object, obligatory adverbial or nothing at all.

It is neither prohibited nor unusual that the UD terminology diverges from the "traditional" one. Partly because there are many traditions, inconsistent with each other. However, it would be nice to at least preserve the distinctions expected in traditional grammar, and to be able to map the UD data to whatever annotation is expected by various communities. Even if UD does not aim at distinguishing arguments from adjuncts universally, the distinction is obviously important in grammars of many languages and there should be standardized means to capture it on the language-particular level.

6 Current UD Annotation

Let us now examine how the core-oblique distinction is dealt with in the current release (2.0) of Universal Dependencies. In order to stay focused on the issues discussed in the previous sections, we limit ourselves to Indo-European languages with case morphology. Table 1 gives an overview. In total, there are 26 UD treebanks (19 languages). Verb-dependent nominals in the data take from 3 to 8 different case forms (including the vocative, which marks a special type of dependent); some nominals are "caseless" (meaning that their annotation does not include the case feature, i.e. either the word does not inflect, or the annotation is incomplete).

Bulgarian and Portuguese represent a larger group of languages where the case system has been reduced to personal pronouns; but only in these two languages the actual numbers for each case surpassed 0.5% of examined nodes. Otherwise, there are all Balto-Slavic languages, all classical Indo-European languages (Ancient Greek, Latin, Gothic, Sanskrit), Modern Greek and German. Some languages have two or three treebanks provided by different groups. Case distribution differs across these treebank sets, but the difference is usually not dramatic. The largest gap can be observed between la_2 and the other two Latin treebanks; besides domain differences, the likely reason is that la and la_3 contain classical Latin while la_2 is from the 13th century.

The differences are more significant when we investigate for each case form whether and how often it occurs with a core dependency relation. Bare nominatives and accusatives are almost always core arguments. Bare datives and genitives also appear as core arguments in convincing numbers. Then the coding seems to be more and more oblique across the ablative, instrumental and locative down to prepositions. Most treebank providers seem to have simply adopted the English rule that oblique are those arguments with prepositions. Occurrences of the `obl` relation among bare noun phrases might as well just mean that these phrases are adjuncts; however, since UD does not distinguish oblique arguments from adjuncts, we cannot verify this hypothesis.

Table 2 is a zoom-in view of cases vs. relations in UD Czech 2.0. The annotation is ported from the Prague Dependency Treebank, which uses the traditional definition where object = argument; that is why the core relations appear in all nominal forms including those with prepositions.

Tables 3 and 4 demonstrate that while the current UD Russian SynTagRus incorporates the En-

glish rule for obliqueness, the first release (1.3), directly converted from the original SynTagRus annotation, was much closer to what we see in Czech. In the 1.3 release, `nmod` under verbs (now labeled `obl`), marked only nominals that are not traditional objects, i.e. adjuncts. In 2.0, these can be no longer distinguished from prepositional objects. Even if it is correct to assume that prepositional arguments are oblique in Russian, there is arguably a substantial amount of information that is important in Russian grammar and was available in the original data, but it is lost in the current UD release.

7 Refined Definition of Objects

Let us now summarize the issues identified in the preceding sections and propose refined guidelines that will hopefully address the issues better (at least in the studied subset of Indo-European languages).

There are three groups of arguments that are traditionally called objects and could be considered as object candidates in UD, ordered by decreasing strength of evidence of their coreness: bare accusatives, bare non-accusatives and prepositional phrases. UD assumes the core-oblique boundary to be clear-cut but it isn't, because identification of primary transitive verbs is not always trivial, and their distribution among the above groups is unbalanced. Nevertheless, drawing the line between bare nominals and prepositional phrases (which is what the majority of treebanks already adopts) seems a reasonable compromise.

In order to preserve the important distinction between prepositional objects and adjuncts, we propose to annotate prepositional objects by the language-specific relation `obl:arg` (except for demoted subjects in passive constructions, which should use `obl:agent`, a practice already established in several UD treebanks).

Bare non-accusatives can be considered core arguments in languages where there are reasonable examples of primary transitive verbs using these cases. (We have shown examples from German and Czech but we have not proved that all cases in all languages from Table 1 meet the criteria. We do believe though that the criteria are met for dative, genitive and instrumental in Slavic languages.) It might be useful to mark them by a language-specific label `obj:nacc`, although it would be just a shortcut: one can obtain the case information from the morphological features.

As for indirect objects, their current UD definition is problematic. It seems appealing to define them as core arguments that are mostly object-like, but grammatical rules applying to them are somewhat different from those used with the *prevailing* type of objects (i.e. the type that covers the largest group of primary transitive verbs). That is, instead of `obj:nacc` proposed above, we would use `iobj` for non-accusative objects (cf. (Karlík et al., 2016)). However, it would also wipe out indirect objects from English, which is a bit unfortunate, given that English seems to be responsible for introducing the very concept of `iobj` in UD. Hence the new guideline should perhaps provide more freedom for language-specific rules, saying that it is possible to mark a subclass of objects as secondary/indirect on language-specific grounds. In the long term, the relation should probably become a language-specific subtype of `obj`.

8 A Note on Subjects

In comparison to the various types of objects, identifying nominal subjects is relatively straightforward in our group of languages. They can be easily recognized by the nominative case and by cross-referencing on the verb (person, number and gender); they can hardly ever be confused with adjuncts. Occasional confusion with objects may stem from morphological ambiguity: in the Czech sentence *Krávy štípou mouchy*, both the nouns *krávy* "cows" and *mouchy* "flies" are in a form shared by nominative and accusative; the (probable) English meaning is "Flies sting cows" but since word order is flexible in Czech, it could also mean "Cows sting flies."

Tables 2 to 4 reveal that a significant subset of subjects in Slavic languages have a genitive form. However, these genitives are caused by numerals in quantified phrases, not by the verb. Under certain conditions, Slavic numerals and quantifiers require that the counted noun takes the genitive form.[7] The numeral itself has its nominative/accusative form, and the entire phrase (numeral + noun) behaves like nominative/accusative singular neuter (gender and number are cross-referenced on the verb). Hence in *Přišlo jen pět dětí* "Only five children came," the verb *přišlo* "came" has a singular neuter form, the numeral

[7] In addition, the genitive can be used partitively without an overt quantifier. In this case it no longer looks like a quantified phrase but it could be understood as one with an elided quantifier.

	nsubj	nsubj:pass	obj	iobj	expl:pv	expl:pass	obl	discourse
Nom (29%)	95	4						
Acc (29%)			69		21	7	2	
Dat (5%)	1		36	33	15		14	1
Gen (3%)	23	1	60	2			14	
Ins (3%)			26	4			69	
Acc+ADP (5%)			37	9			54	
Dat+ADP (2%)			31	3			66	
Gen+ADP (7%)	1		7	2			89	
Loc+ADP (12%)			10	2			88	
Ins+ADP (3%)			28	5			66	
None (2%)	58	1	19	6			12	
None+ADP (1%)	1		13	3			83	

Table 2: UD Czech. Distribution of core and oblique relations for individual case forms. Numbers indicate how many % of the nominals in the given case got the given relation. ADP indicates a preposition.

	nsubj	nsubjpass	dobj	iobj	nmod	nmod:agent
Nom (27%)	85	13	1			
Acc (17%)			97		2	
Dat (2%)			36		64	
Gen (4%)	30	4	51	2	13	
Ins (5%)			32		51	16
Acc+ADP (7%)			31		69	
Dat+ADP (3%)			29		70	
Gen+ADP (7%)	1		24	17	57	
Loc+ADP (11%)					98	
Ins+ADP (3%)			27		73	
None (12%)	60	3	26		10	1
None+ADP (2%)			30	7	63	

Table 3: UD Russian SynTagRus 1.3. Distribution of core and oblique relations for individual case forms. Numbers indicate how many % of the nominals in the given case got the given relation. ADP indicates a preposition.

	nsubj	nsubj:pass	obj	iobj	obl	obl:agent
Nom (34%)	90	8			1	
Acc (20%)			97		2	
Dat (3%)				5	95	
Gen (5%)	27	2		1	70	
Ins (6%)					79	17
Acc+ADP (7%)					99	
Dat+ADP (3%)					99	
Gen+ADP (7%)	1			6	93	
Loc+ADP (11%)					99	
Ins+ADP (3%)					100	

Table 4: UD Russian SynTagRus 2.0. Distribution of core and oblique relations for individual case forms. Numbers indicate how many % of the nominals in the given case got the given relation. ADP indicates a preposition.

pět "five" is in nominative and the noun *dětí* "children" is in genitive. Counted phrases are headed by nouns in UD, thus the genitive noun is attached directly to the verb; but a language-specific relation between the noun and the numeral preserves the information about who governs the case.

It has also been discussed[8] whether certain constructions in Slavic languages sanction subjects in the dative. An example (Russian) is *Мне было холодно* / *Mne bylo holodno* lit. "To-me it-was cold," meaning "I was cold." The dative argument *мне* is called *logical subject* by some grammarians. However, under the UD guidelines it will be subject only if it receives the treatment normally accorded to the single argument of a one-argument predicate in Russian. This "normal treatment" includes nominative case marking, but not only that. Its gender and number should be cross-referenced on the predicate, but *было холодно* is neuter singular regardless of the referent of *мне*. And finally, if the clause is converted to infinitive and complements another predicate, the infinitive should inherit the subject from the matrix clause. However, the dative pronoun cannot be removed and make room for an inherited subject. We still have it in "he will stop to be cold": *ему перестанет быть холодно* / *emu perestanet byt' holodno*. The verb "to stop" takes a normal nominative subject but if we provide it, the sentence becomes ungrammatical: **он перестанет быть холодно*. Thus the dative argument failed on all three accounts; on the other hand, the treatment it receives is not unlike the dative objects in Russian. Note that we are not saying that all subjects in all Indo-European languages must be nominative.[9] The point is that there usually is some typical treatment of subjects in the given language; the said dative argument does not receive the treatment typical in Russian, thus it is not subject.

9 Conclusion

We have reviewed the methodology proposed by (Andrews, 2007) for distinguishing core/oblique arguments; in particular, we have shown how it applies to the case morphology observed in a number of Indo-European languages. While UD focuses on core arguments in order to avoid distinguishing arguments from adjuncts, we observe that the distinction is needed (to some extent) to recognize core arguments. Similarly, UD does not label semantic roles but we still must consider them in order to recognize primary transitive verbs. Overall we found the method very useful (actually the only practically usable approach that has been proposed so far in the context of UD) but it has to be applied carefully and it does not provide absolute criteria (probably nothing does). If the properties of core arguments in all UD languages are defined following the principles we showed for German, Czech and Russian, the UD annotation will become much more consistent cross-linguistically than it is now.

We have also shown that defining objects in terms of core arguments conflicts with the traditional view in some languages, where all arguments are objects. We do not want to reject the core-oblique perspective; nevertheless, we propose to use the `obl:arg` relation and preserve the argument-adjunct distinction in UD if it is available.

[8] `http://github.com/UniversalDependencies/docs/issues/248`

[9] In fact, (Andrews, 2007) gives an example of a dative subject in Icelandic.

Acknowledgments

The work was supported by the grant 15-10472S of the Czech Science Foundation.

References

Avery D. Andrews. 2007. The major functions of the noun phrase. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure. Second edition, pages 132–223. Cambridge University Press, Cambridge, UK.

Robert M. W. Dixon. 2012. Basic Linguistic Theory. Oxford University Press, Oxford, UK.

Matthew S. Dryer. 2007. Clause types. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure, pages 224–275. Cambridge University Press, Cambridge, UK.

William A. Foley. 2007. A typology of information packaging in the clause. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure, pages 362–446. Cambridge University Press, Cambridge, UK.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková-Razímová. 2006. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Bohuslav Havránek and Alois Jedlička. 1966. Stručná mluvnice česká. Fortuna, Praha, Czechia.

Gerhard Helbig and Joachim Buscha. 1998. Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht (18. Auflage). Langenscheidt, Leipzig, Germany.

Petr Karlík, Marek Nekula, Jana Pleskalová, et al. 2016. Nový encyklopedický slovník češtiny. Nakladatelství Lidové noviny, Praha, Czechia.

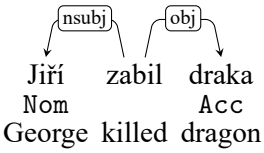
Markéta Lopatková, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. 2016. Valenční slovník českých sloves VALLEX. Karolinum, Praha, Czechia.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

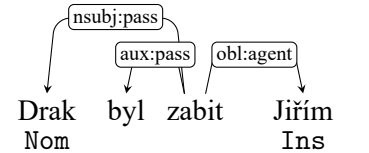
Roland Sussex and Paul Cubberley. 2006. The Slavic Languages. Cambridge University Press, Cambridge, UK.

Appendix A. Czech Examples

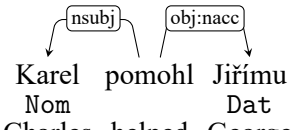
(1)



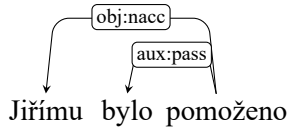
(2)



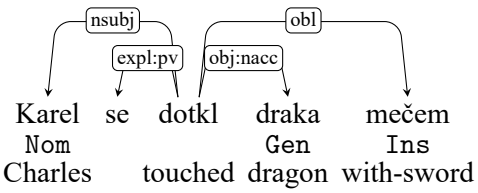
(3)



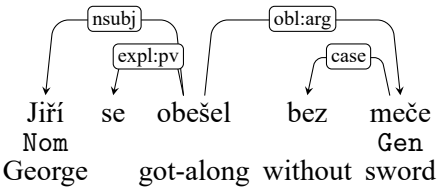
(4)



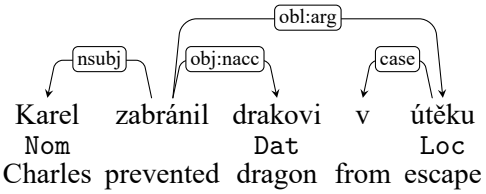
(5)



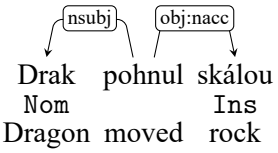
(6)



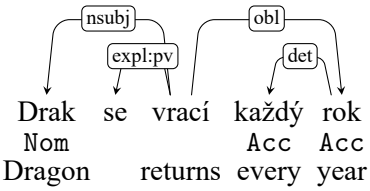
(7)



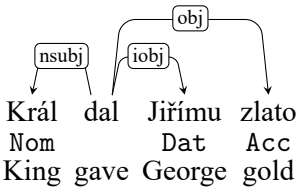
(8)



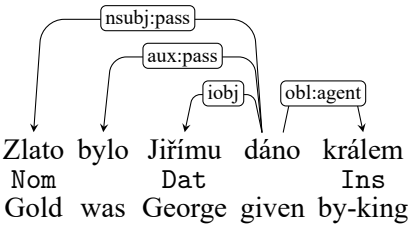
(9)



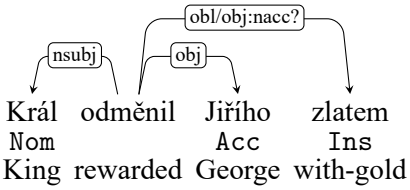
(10)



(11)



(12)



(13)