Towards a Unified Taxonomy of Deep Syntactic Relations

Kira Droganova, Daniel Zeman
Charles University, Faculty of Mathematics and Physics, ÚFAL
Prague, Czechia
{droganova, zeman}@ufal.mff.cuni.cz

Abstract

This paper analyzes multiple deep-syntactic frameworks with the goal of creating a proposal for a set of universal semantic role labels. The proposal examines various theoretic linguistic perspectives and focuses on Meaning-Text Theory and Functional Generative Description frameworks and PropBank. The research is based on the data from four Indo-European and one Uralic language – Spanish and Catalan (Taulé et al., 2011), Czech (Hajič et al., 2017), English (Hajič et al., 2012), and Finnish (Haverinen et al., 2015). Updated datasets with the new universal semantic role labels are now publicly available as a result of our work. Nevertheless, our proposal is oriented towards Universal Dependencies (UD) (de Marneffe et al., 2021) and our ultimate goal is to apply a subset of the universal labels to the full UD data.

Keywords: Semantics, Deep representation, Language resource

1. Introduction

Linguistic research and multilingual natural language processing need annotated data in many languages, ideally following a uniform annotation framework. For morphology and surface syntax, Universal Dependencies (UD)[1] (de Marneffe et al., 2021) is the current de-facto standard of such a framework. Nevertheless, despite being an important linguistic resource, UD is only one step towards natural language understanding. The mapping between surface syntax and meaning is not straightforward, as the same meaning can be encoded in various syntactic constructions (e.g., active vs. passive clauses), and vice versa, one syntactic construction can be used to convey different meanings (e.g., the English preposition *on* can express location, time, or other verb-specific roles as in *I rely on him*). Therefore there are datasets that attempt to annotate another layer (or multiple layers) of the language, which is closer to the meaning and is variously termed 'deep-syntactic', 'tectogrammatical', or even 'semantic'. Unfortunately, the annotations in this layer have not reached the level of cross-linguistic uniformity and interoperability as UD set for morphology and surface syntax.

Deep-syntactic annotation can cover a variety of phenomena but in the present paper, we focus on the inventory of deep-syntactic (or semantic) relations between words. We have selected the approaches that have been extensively studied for a longer period of time, and have been utilized in natural language applications, such as Meaning-Text Theory (Kahane, 2003), Functional Generative Description frameworks (Sgall, 1967) and PropBank (Kingsbury and Palmer, 2002). We study the inventories used in these frameworks, compare them and propose a unified inventory where the same meaning would have the same label across datasets. This unified set of relations should be applicable to any language. Ideally, it should be possible to map relations from existing frameworks onto this inventory without loss of information; while there is no guarantee that this ideal goal is achievable, we want to get as close to it as possible.

There are two related projects worth mentioning here. Universal Proposition Bank (Jindal et al., 2022) provides semantic role annotation for 23 languages, based on their UD treebanks. As the name suggests, semantic role labels follow the PropBank (Kingsbury and Palmer, 2002). Second, a recent proposal by Evang (2023) defines the CRANS annotation scheme in order to annotate semantic roles on top of UD. Only a few coarse and cros-linguistically applicable valency frames (superframes) are defined in CRANS in order to avoid reliance on large-coverage, language-specific valency dictionaries.

We first survey the deep-syntactic relations in Meaning-Text Theory (Section 2), Functional Generative Description (Section 3), and PropBank (Section 4).[2] We provide a comparison of the

[1] https://universaldependencies.org/

[2] Our approach was rather opportunistic: We were able to find data for the selected frameworks. This is also the reason why our current language pool is not too varied typologically (there are four Indo-European and one Uralic language). One could ask, for example, what would happen if we worked with an ergative language. We assume that the necessary inventory of semantic roles will not change much (if at all), but their mapping to syntax can be quite different of course.

LREC-COLING 2024, pages 16412–16421
20-25 May, 2024.

frameworks (Section 5). Finally, in Section 6 we propose a unified set of relations to which the other three can be mapped.

2. Meaning-Text Theory

2.1. Overview

The goal of the Meaning-Text Theory (MTT) is to write systems of explicit rules that express the correspondence between meaning and text (or sound) in various languages (Kahane, 2003). MTT defines a seven-level representation that describes the relation between form and meaning.
The set of deep-syntactic relations used in MTT consists of numbered arguments and "utility" relations such as **ATTR** for attributes and other modifiers, **COORD** for coordination, and **APPEND** for parentheses, interjections, and other similar items.

2.2. Thematic Roles

The AnCora corpus of Catalan and Spanish (Taulé et al., 2008)[3] was used to examine the set of semantic relations defined in MTT.
Deep syntactic / semantic relations are assigned to up to seven argument slots (**arg0**, **arg1**, **arg2**, **arg3**, **arg4**, **argM** and **argL**) and 20 thematic roles. Each of the roles can be mapped to several syntactic functions and argument positions. The arguments required by the verb sense are incrementally numbered, expressing their degree of proximity in relation to its predicate (Palmer et al., 2005). The two unnumbered argument slots are **argM** for adjuncts and **argL** for lexicalized complements of light verbs.

2.2.1. Adverbial: adv

The Adverbial role is a broad category that corresponds to non-specific adjuncts and can be expressed by the UD syntactic relations `advcl`, `advmod` or `obl`.

2.2.2. Agent: agt

The Agent role is associated with the external causer argument that is expressed as the syntactic subject. In some cases the external argument (**arg0**) may be expressed as an oblique agent complement, keeping its original Agent role as well. The Agent role can be expressed syntactically as `nsubj`, `det`, `nmod`, and `obl` as in *El gol fue convertido* ***por Rodrigo Barra*** "The goal was scored **by Rodrigo Barra**".

2.2.3. Attribute: atr

The Attribute role refers to the third position (**arg2**). It is typically expressed as the direct object. Other examples that can be found in the data are `ccomp` and `root` (Figure 1).

[3]The two corpora were converted to dependencies (Hajič et al., 2009) and later to Universal Dependencies. UD version 2.12 was used for this paper.

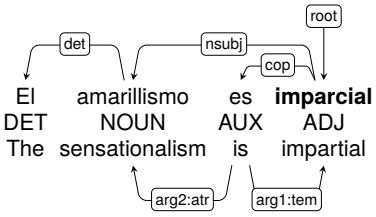

Figure 1: *The sensationalism is impartial.* An example of **Arg2:atr** – root. In the original AnCora the adjective *imparcial* "impartial" is analyzed as an argument of the copula. In UD the copula is treated as a functional attribute of the adjective, while the adjective becomes the predicative root of the sentence.

2.2.4. Beneficiary: ben

The Beneficiary role refers to the third argument (**arg2**). In UD it is expressed as `obl` as in *Esto permitirá* ***al banco*** *sanear su portafolio* "This will allow **the bank** to clean up its portfolio".

2.2.5. Cause: cau

The Cause role corresponds to the external causer argument that is typically the syntactic subject. The Cause role can also take an adjunct position. In that case it receives the UD labels `obl` or `advcl` as in *Me gusta este trabajo,* ***porque aquí hay mucho por hacer*** "I like this job **because there is a lot to do here**".

2.2.6. Cotheme: cot

The Cotheme role refers to the third argument position (**arg2**). This role is expressed as a prepositional object – the UD labels `nmod` or `obl` as in *El viaducto peatonal conecta Mollet* ***con Martorell*** "The pedestrian viaduct connects Mollet **with Martorell**".

2.2.7. Destination: des

The Destination role typically corresponds to the fifth argument position (**arg4**) that is most frequently expressed as `obl` and `nmod`.

2.2.8. Experiencer: exp

The Experiencer role refers to the first argument (**arg0**) that is expressed as the subject. In some cases it can correspond to the third argument (**arg2**) that is expressed as the UD label `obl`.

2.2.9. Final State: efi

The Final State role refers to the third argument position (**arg2**). It can be expressed as a predicative complement or a prepositional object — the UD label `obl` as in *Ni siquiera ha llegado* ***a sietemesino*** "Not even has reached **seven months**".

2.2.10. Initial State: ein

The Initial State role is similar to the Final State role with the difference that it occurs in the data

less frequently. It refers to the third argument position (**arg2**) and can be expressed as a prepositional object or a predicative complement as in *El que la derriba, ha ido **de la insatisfacción** a la violencia.* "The one who tears it down has gone **from dissatisfaction** to violence."

2.2.11. Instrument: ins

The Instrument role refers to the third argument position (**arg2**) that is typically expressed as `obl` in UD as in *Los policías, equipados **con material antidisturbios**, se mantendrán atentos a posibles incidentes* "The police, equipped **with riot gear**, will remain alert to possible incidents".

2.2.12. Location: loc

The Location role can be expressed as the third argument (**arg2**) marked with the UD labels `obl` or `obl:arg`.

2.2.13. Manner: mnr

The Manner role refers to an adjunct (**argM**) that can receive one of the following syntactic labels: `obl` or `advmod`.

2.2.14. Origin: ori

The Origin role occurs in the data less frequently; It marks the place of origin and typically takes the fourth argument position (**arg3**). The most frequent syntactic label is `obl`.

2.2.15. Patient: pat

The Patient refers to the second argument position (**arg1**) that is typically expressed as the direct object. It can also be expressed as the syntactic subject as in ***Cualquiera** que lo hiciese con más jugadores relevándolos habitualmente era considerado como un loco* "**Anyone** who did it with more players relieving them was usually considered crazy".

2.2.16. Purpose: fin

The Purpose role refers to an adjunct; most frequently it is expressed as `advcl` on the syntactic level as in ***Para entendernos**, diríamos esperpentos* "**To understand each other**, we would say grotesque"

2.2.17. Source: src

The Source role refers to the first argument position (**arg0**) represented by the UD label `nmod` as in *La catedral padeció una oleada **de pintadas*** "The cathedral suffered a wave of graffiti".

2.2.18. Theme: tem

The Theme role typically takes the second argument position (**arg1**). Most frequently it receives one of the following syntactic labels: `nsubj`, `obj`, `nmod`, and `obl`.

2.2.19. Time: tmp

The Time role refers to temporal adjuncts that most frequently receive the following syntactic labels: `obl`, `advmod`, and `advcl`.

2.2.20. Empty label: argL

The **argL** slot refers to the lexicalized arguments of light verbs. This slot does not receive any role label and most frequently occurs as `obl` or `obj` (Figure 2) on the syntactic level.

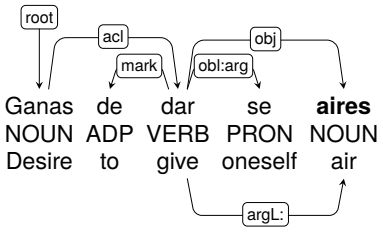

Figure 2: *The desire to show off.* An example of **ArgL:** – `obj`. The idiomatic light verb construction *darse aires* (lit. *give oneself air*) means "to brag, to show off".

3. Functional Generative Description

3.1. Overview

Functional Generative Description (FGD) represents a dependency-based generative description that is based on a multilayer design reflecting the relation of form and function (Sgall, 1967).

Deep syntactic information is captured by the tectogrammatical representation that describes the meaning of the sentence. Thus synonymous sentences have a single representation at this level, while an ambiguous sentence has more than one tectogrammatical representation.

3.2. Semantic Role Labels

FGD serves as a basis for the Prague Dependency Treebank (Hajič et al., 2006; Bejček et al., 2013) and its successors such as Prague Czech-English Dependency Treebank (Hajič et al., 2012). The original semantic role labels (functors) have been carried over to the UD data.[4] The same conversion was applied to the Prague Czech-English Dependency Treebank (PCEDT).[5]

There are 67 semantic roles (functors), divided into arguments (actants, inner participants) and adjuncts according to both semantic and formal criteria specified within the valency theory (Panevová, 1974).

[4]Available since UD v2.12.

[5]Access to the UD version of PCEDT is restricted due to license terms. The original PCEDT is available through the Linguistic Data Consortium.

3.2.1. Argument Functors

FGD specifies five argument roles that correspond mostly to the core arguments (subject, direct and indirect object) of the verb on the surface-syntactic layer.

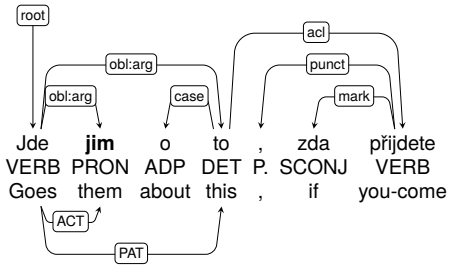

Figure 3: *Their point is whether you are coming.* A non-canonical example where **ACT** is a dative (oblique) argument.

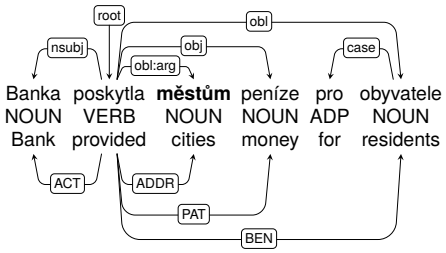

Figure 4: *The bank provided cities with money for residents.* An example of **ADDR** realized as dative oblique argument, and **BEN** realized as prepositional oblique dependent.

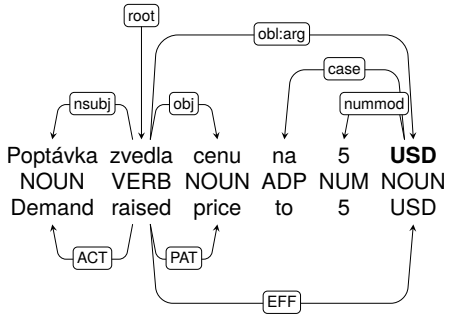

Figure 5: *The demand has pushed the price to $5.* An example of **EFF** realized as an oblique argument.

ACT argument: actor; mostly `nsubj`, sometimes `obl:arg` (Figure 3), `obl:agent` (in passive clauses), or even `obl`.
PAT argument: patient; mostly `obj`, also `nsubj:pass`, or even `nsubj` as in *Bolí ho* ***noha*** "His **leg** hurts".
ADDR argument: addressee; mostly `obl:arg` (Figure 4).

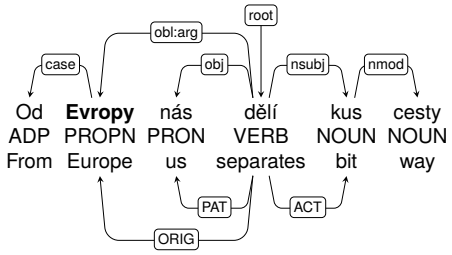

Figure 6: *A long way separates us from Europe.* An example of **ORIG** realized as an oblique argument.

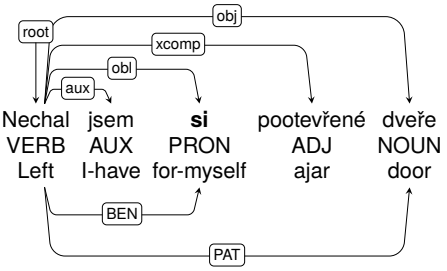

Figure 7: *I left the door ajar for myself.* An example of **BEN** realized as dative oblique dependent.

ORIG argument: origo; mostly `obl:arg` (Figure 6).
EFF argument: effect; mostly `ccomp`, `obl:arg` (Figure 5).

Other types of verbal modifications are considered adjuncts. These functors correspond to temporal, locational, manner and other kinds of adverbials. They can be classified by their intended purpose and typically occur as **obl**, **advmod**, and **nmod** in UD.

3.2.2. Locative and Directional Adjuncts

Locative and directional functors express location or direction related to the content of the governing word (see Table 4 for details).

3.2.3. Temporal Adjuncts

Temporal functors express various temporal points or intervals. Individual temporal functors differ according to which of the possible questions about time they answer (see Table 5 for details).

3.2.4. Manner Adjuncts

Functors for expressing manner constitute a broad category of adjuncts that express the inner characteristics of events such as comparison, specifying the result of an event or the manner an action is performed (see Table 2 for details).

3.2.5. Causal Adjuncts

Functors for causal relations express various causal relations between events or states such as cause, condition, purpose, or concession (see Table 3 for details).

3.2.6. Specific Adjuncts

The following functors are assigned to certain specific modifications that are not traditionally included in the syntactic descriptions. They are close to manner adjuncts (see Table 2 for details).

3.2.7. Adnominal Functors

Specific adnominal functors are designed exclusively for modifying (semantic) nouns (see Table 8 for details).

3.2.8. Rhematizer Functors

Functors for rhematizers, sentence, linking and modal adverbial expressions are designed for representing free modifications and their function in the sentence – to rhematize, to link the sentence to its preceding context or to express various modal meanings and attitude (see Table 9 for details).

3.2.9. Functors for Multi-word Lexical Units

This group of functors is used for representing certain multi-word lexical units or foreign-language parts that are not strictly analyzed (see Table 9 for details).

3.2.10. Paratactic Structures

This group of functors expresses the relations between the members of paratactic structures (either clauses or modifications) (see Table 7 for details).

3.2.11. Independent Clauses

Functors for the effective roots of independent clauses express the independence of the given lexical unit and determine the clause type (see Table 6 for details).

3.2.12. Other Functors

The **COMPL** functor is assigned to predicative complements.

The **CM** functor is assigned to conjunction modifiers, mostly various particles and adverbs.

4. PropBank

4.1. Overview

The Proposition Bank project focuses on the argument structure of verbs, including roles traditionally viewed as either arguments or adjuncts, adding a layer of predicate-argument information, or semantic role labels, to the constituent structures in the Penn Treebank (Palmer et al., 2005).

4.2. Semantic Labels

To broaden the typological variability of our research, we have included (in addition to English) the Finnish Proposition Bank (Haverinen et al., 2015). It is built on top of the Turku Dependency Treebank.[6]

The English PropBank frames were propagated from the Penn Treebank (Marcus et al., 1993) constituent trees to their UD conversion.[7]

The PropBank defines semantic roles on a verb-by-verb basis. An individual verb's semantic arguments are numbered, beginning with 0 and ending with 5. **ARG0** is generally the argument exhibiting features of a Prototypical Agent (Dowty, 1991), while **ARG1** is a Prototypical Patient or Theme. No consistent generalizations can be made across verbs for the higher-numbered arguments, though an effort has been made to consistently define roles across members of VerbNet classes (Palmer et al., 2005).

- **ARG0**: mostly `nsubj`
- **ARG1**: mostly `obj`, `nsubj`; in English also `ccomp`: *Mr. Spielvogel said* ***he hopes that …***
- **ARG2**: mostly `obl`, `root`[8] in the Finnish data; in English also `obj`
- **ARG3**, **ARG4**, **ARG5**: mostly `obl`; in Finnish also `advmod`
- **ARGA**: used for a causal agent, only with verbs of volitional motion as in ***He*** *followed offers with threats*.
- **ARGM**: used for adjuncts, whose meaning is not tightly bound to a particular verb.

PropBank also uses a number of subtypes that further specify the function or semantic category of a participant. They can extend the main label, as in **ARGM-LOC**. They are most often seen with adjuncts (**ARGM**), but they can also extend the numbered arguments, and some are even more frequent with numbered arguments than with adjuncts (e.g., **ARG1-PRD** or **ARG2-EXT**). The Finnish data differs from English more in the usage of subtypes than in other aspects.

- **LOC**: location; mostly `obl`
- **DIR**: direction; mostly `obl`, in English also `compound`, `case`
- **TMP**: time; mostly `obl`, `advmod`
- **MNR**: manner; mostly `advmod`, `obl`
- **CAU**: cause; mostly `advcl`, `obl`
- **PNC**: purpose;[9] mostly `obl`, `advcl`
- **ADV**: other adverbial; mostly `advmod`, `obl`
- **EXT**: extent; mostly `obl`, in Finnish also `advmod`
- **PRD**: for secondary predication; mostly `obl`, in English also `xcomp`

[6] For compatibility reasons, the treebank was converted from UD v1 to UD v2 using UDPipe 2 (Straka et al., 2021) trained on UD 2.12, available at https://lindat.mff.cuni.cz/services/udpipe/api-reference.php.

[7] UD conversion of PCEDT was used.

[8] In copula constructions.

[9] Purpose clauses are marked with **PRP** in the English data and with **PNC** in the Finnish data.

- **REC**: reciprocal; mostly `advmod`
- **MOD**: modal verb; mostly `aux`
- **NEG**: negation marker; mostly `aux` in the Finnish data, and `advmod` in English
- **DIS**: discourse connectives, such as *still, also, however, but, for example*; mostly `advmod`, in Finnish also `obl`
- **CSQ**: consequence; only Finnish, mostly `acl:relcl`
- **PRT**: phrasal marker; only Finnish, mostly `obl`, `compound`

5. Comparison of the Frameworks

We will now present some observations on the similarities and divergencies across the three frameworks introduced above. The comparison will serve as the basis for our proposal of a unified taxonomy in Section 6.

PDT defines functors for all content words in the sentence, including modifiers of nominals. AnCora and PropBank define them only for participants of events; they can be nominals and not necessarily verbs, but they must denote events.

5.1. Inner Participants

We do not have space to assess all labels here due to their high number, especially on the PDT side. Therefore we will concentrate on the most important ones, namely the five arguments ('inner participants').

Actor, agent The PDT label **ACT** and the AnCora label **arg0:agt** have very similar meanings but they are not identical. **ACT** is defined relatively as the most active participant; it does not have to be too active if there are no other participants. So it includes experiencers, which have a separate label in AnCora: **arg0:exp**. It even includes inanimate subjects in change-state clauses (***Las reservas de oro*** *subieron 800 millones de dólares* "**Gold reserves** rose by $800 million"), which AnCora labels as themes (**arg1:tem**). Causer in causative constructions is **arg0:cau** in AnCora but **ACT** in PDT (but note that there are also subordinate clauses of cause, which are **argM:cau** in AnCora and **CAUS** in PDT).

In PropBank, **ARG0** typically corresponds to AnCora's **arg0:agt**. Similar reservations hold between PropBank and PDT as between AnCora and PDT. In ***The door*** *opened*, *the door* is **ACT** in PDT but **ARG1** in PropBank because the opening was caused by some unexpressed agent. This contrasts with the passive sentence ***The door*** *was opened*, where both PDT and PropBank regard the subject as **PAT** / **ARG1**. In certain causative frames, PropBank's **ARGA** is AnCora's **arg0:cau**.

Patient The PDT label **PAT** is the participant affected by the action, and the second most prominent one; if a predicate has only two arguments, the less active of the two will be **PAT**. It is most likely to correspond to **arg1:pat** in AnCora; however, with predicates where the object is more abstract, it may be labeled **arg1:tem** (*cubrir* ***empleos en diversos sectores*** "to fill jobs in various sectors").

In the majority of cases the PropBank label **ARG1** corresponds to the PDT label **PAT**; but sometimes it maps to **ACT** (see above) or to **EFF** (see below). PDT **PAT** is also used for secondary predication, as in *His client contacts could prove* ***a gold mine***; such cases are labeled **ARG2-PRD** in PropBank.

Addressee The PDT label **ADDR** is the addressee or recipient in events of giving and transfer. In these events, the addressee slot is licensed by the verb and is thus treated as inner participant. In contrast, almost every event can have an adjunct expressing beneficiary. It gets a different label, **BEN** (Figures 4 and 7). AnCora does not seem to make the distinction. It uses **arg2:ben** in both cases.

The PropBank label **ARG2** typically corresponds to the PDT label **ADDR**, as in ***Sallalle*** *osaan kyllä neuleen kuvitella* "**For Salla**, I can certainly imagine the sweater." **ARG2-EXT** primarily corresponds to the broader PDT labels of **DIFF**, **EXT**, and **REG**, which express manner by specifying differences between the compared events, by indicating the extent or intensity of the event, and by saying with respect to what something holds. Furthermore, we mentioned above that **ARG2-PRD** would be **PAT** in PDT.

Effect The PDT label **EFF** denotes the result of an event. While it is licensed by the verb and thus inner participant, it is usually not obligatory. A frequent example is changed value, as in *Poptávka zvedla ceny stříbra* ***až na 5,5 dolarů za trojskou unci*** "The demand pushed silver prices **up to $5.50 per troy ounce**" (Figure 5); similar examples in AnCora are labeled **arg4:des** (destination). On the other hand, the destination role in AnCora also applies to directional adjuncts in movement events, which would be labeled **DIR3** in PDT. PDT also uses **EFF** in events that are not changes of state in the true sense: *být považován* ***za vůdčí osobnost*** "to be considered **a leading figure**". In AnCora, such an argument would be **arg2:atr** (attribute).

While any generalization about PropBank's **ARG3** to **ARG5** has to be taken with a grain of salt, many instances of **ARG3** and **ARG4** correspond to **EFF** in PDT, e.g., *The union, though, has called the offer* ***"insulting"***. In some cases **ARG1** may correspond to **EFF** as in *GM officials said they, too,* ***were surprised by the move***.

Origo The PDT label **ORIG** is the least frequent of the five argument functors. While it may cor-

respond to the English preposition *from*, it is not used for adjuncts of direction (those would get **DIR1**) or time (those would get **TFRWH** or **TSIN**). Being licensed by the verb, it is used, e.g., for material from which something is made, and also when distinguishing or separating something from something else: ***Od Evropy*** *nás dělí velký kus cesty* "A long way separates us **from Europe**" (Figure 6). Similar examples in AnCora are annotated as **arg2:cot** (cotheme). In contrast, the seemingly corresponding label **arg3:ori** is used for directional and temporal adjuncts.

In PropBank, similar constituents are higher-numbered arguments. A relatively frequent example is **ARG3-from** as in *to change its emphasis* ***from buying mortgage loans***.

5.2. Adjuncts and PropBank-specific Subtypes

About a half of the PDT functors (32) classify adverbial adjuncts. In AnCora and PropBank, adjuncts are the subtypes of **argM** / **ARGM**. Many of them are self-explanatory and can be mapped easily, the only problem being different levels of granularity in the three frameworks. We highlight some interesting cases here. Recall that PropBank subtypes, although generally intended for **ARGM**, sometimes occur with numbered arguments and some of them are even more likely to accompany a numbered argument than **ARGM**.

ARGM-EXT usually corresponds to the PDT label **PAT** as in *Aetna closed* ***at $60***.

The PropBank label **PRD** is designed for secondary predication and typically used as a subtype of arguments (rather than adjuncts). It corresponds to multiple argument labels in PDT such as **PAT** as in *Many of the morning-session winners turned out* ***to be losers*** *by afternoon*, and **EFF** as in *It was just the culture of the industry that kept it* ***from happening***.

The discourse connective label **ARGM-DIS** within PropBank conforms to the PDT labels **PREC** (a functor linking the clause to the preceding context) and **RHEM**, which represents a rhematizer, including negative expressions. Negative expressions are distinguished by **ARGM-NEG** in PropBank.

The PropBank label **MOD** is used for modal verbs as in *The notes* ***can*** *be redeemed*. It differs from the PDT label **MOD**, which represents modal adverbs or particles: ***Maybe*** *Mr. Z. was too busy*. Modal verbs in PDT are treated as attributes of the main verbs and do not receive their own functors.

The PropBank label **CAU** is used for causative constructions such as *Tuota tekisi mieli sovittaa,* ***sillä vartaloni tarvitsee aivan tietyn mallisen mekon*** "I would like to fit that, **because my body needs a dress of a very specific model**". This label directly corresponds to the **CAUS** label in PDT.

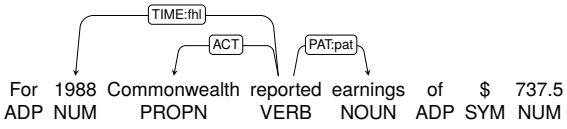

Figure 8: Unified semantic role annotation of the sentence *For 1988 Commonwealth reported earnings of $737.5* .

The purpose clauses labeled as **PNC** in the Finnish data or **PRP** in the English data correspond to PDT functors indicating causal relations. The majority of cases match the PDT label **AIM** as in *Sinne ne sitten jäävät* ***kurottelemaan tavoitteitaan*** "That's where they stay **to reach their goals**". Some cases correspond to **BEN** as in *Mr. Achenbaum will do some strategic consulting at the agency* ***for 'non-clients'*** or **CAUS** as in *Procardia, a heart medicine, have shrunk* ***because of increased competition***.

The PropBank label **REC** is used for reciprocity as in *Both men seemed to work well* ***together***. In PDT, reciprocity is represented on the tectogrammatical layer by inserting a node with the #Rcp t-lemma in the position of the omitted valency slot. This means that there is no separate PDT functor for reciprocals; instead, the node has the functor corresponding to the unoccupied valency position (**PAT** in the majority of cases).

6. Unified Semantic Role Labels

Our goal is to define a label set capable of capturing as many distinctions from the source frameworks as possible. The role classification has to be hierarchical so that less granular source labels can be mapped to less specific labels (higher level in the hierarchy). The proposed set consists of 13 top-level labels. A unified semantic role label is structured as follows: *MAIN:subcategory*. Figure 8 shows an example sentence annotated with unified semantic role labels.

The *MAIN* label expresses the main semantic category. Some FGD-motivated labels step out of the line and classify paratactic relations or other phenomena. The subcategory part is optional. It is designed for preserving finer distinctions that are available in some frameworks. For instance, the LOC label (Table 4) is intended for denoting location or direction. P(CE)DT and AnCora have the capability to distinguish between specific categories within the direction category. In this case, this information can be preserved and a sub label can be assigned (LOC:dir1, LOC:dir2 or LOC:dir3). PropBank has a single label for direction, so only the coarse grained LOC label can be assigned.

As a practical result, the datasets for the five languages discussed in the paper is accessible at `http://hdl.handle.net/11234/1-5474`.

The datasets provide unified labels to enhance interoperability and support cross-lingual studies.

6.1. Arguments

Although the approach does not lean towards a valency dictionary, we loosely follow the PDT distinction between arguments and adjuncts (Table 1) keeping in mind that it will not be precise due to varying definitions of arguments and adjuncts across frameworks.

Label	Sublabel	P(CE)DT	AnCora	PropBank
ACT	ACT:agt ACT:exp ACT:cau	ACT: actor	agt exp arg0:cau	ARG0 ARGA
PAT	PAT:pat PAT:theme PAT:atr	PAT: patient	pat arg1:tem atr	ARG1 ARGM-REC ARG1
ADDR	ADDR	ADDR: addressee	ben	ARG2
EFF	EFF	EFF: effect	efi des src	ARG3 ARG4
ORIG	ORIG	ORIG: origo	cot ein	ARG3-from

Table 1: Arguments

6.2. Manner: MANR

The **MANR** label refers to adjuncts of manner that describe how the action, experience, or process of an event is carried out (Table 2).

Label	Meaning	P(CE)DT	AnCora	PropBank
MANR:mann	manner	MANN	argM:mnr	ARGM-MNR
MANR:means	means	MEANS	argM:ins	ARGM-ADV
MANR:resl	result	RESL	N/A	
MANR:ext	extent	EXT	N/A	ARG2-EXT
MANR:diff	difference	DIFF	N/A	
MANR:cpr	comparison	CPR	N/A	
MANR:contrd	contrast	CONTRD	N/A	
MANR:subs	substitution	SUBS	N/A	
MANR:restr	exception	RESTR	N/A	
MANR:acmp	accompaniment	ACMP	N/A	ARGM-ADV
MANR:ben	benefactor	BEN	argM:ben	
MANR:her	inheritance	HER	N/A	
MANR:crit	criterion	CRIT	N/A	
MANR:reg	regarding	REG	N/A	
MANR:compl	pred. complement	COMPL	N/A	N/A

Table 2: MANR Role

6.3. Causal: CAUSE

The **CAUSE** label refers to adjuncts that express various causal relations (Table 3).

Note that there are two additional causal relations that are paratactic and thus clustered with the **BINDER** roles (Section 6.7): **reas** and **csq**.

Label	Meaning	P(CE)DT	AnCora	PropBank
CAUSE:aim	purpose	AIM	argM:fin	ARGM-PNC/PRP
CAUSE:intt	intention	INTT		ARG2-PRD
CAUSE:caus	cause	CAUS	argM:cau	ARGM-CAU
CAUSE:cncs	concession	CNCS	argM:adv	ARGM-ADV
CAUSE:cond	condition	COND	-	

Table 3: CAUSE Role

6.4. Locative: LOC

The **LOC** label is bound to location or direction (Table 4).

Label	Meaning	P(CE)DT	AnCora	PropBank
LOC:where	where	LOC	argM:loc	ARGM-LOC
LOC:dir1	where from	DIR1	argM:ori	
LOC:dir2	which way	DIR2	argM:loc	ARGM-DIR
LOC:dir3	where to	DIR3	argM:des	

Table 4: LOC Role

6.5. Temporal: TIME

The **TIME** label refers to temporal adjuncts that express various temporal points or intervals.

Label	Meaning	P(CE)DT	AnCora	PropBank
TIME:when	when	TWHEN		
TIME:frwh	from when	TFRWH		
TIME:sin	since when	TSIN		
TIME:par	in parallel with what	TPAR		
TIME:fhl	for how long	TFHL	tmp	ARGM-TMP
TIME:hl	(after) how long	THL		
TIME:towh	to when	TOWH		
TIME:till	until when	TTILL		
TIME:ho	how often	THO		

Table 5: TIME Role

6.6. Independent Clauses: IND

The **IND** label is designed for the functors that express the independence of the given lexical unit and determine the clause type (Table 6).

Label	Meaning	P(CE)DT	AnCora	PropBank
IND:pred	independent clause	PRED		
IND:par	parenthetic clause	PAR		
IND:denom	independent nominal	DENOM	N/A	N/A
IND:vocat	vocative	VOCAT		
IND:partl	independent interjection	PARTL		

Table 6: IND Role

6.7. Paratactic: BINDER

The **BINDER** label refers to paratactic structures and captures the relation between different parts of the utterance (Table 7).

6.8. Adnominal: ADNOM

The **ADNOM** label is designed for modifiers of (semantic) nouns (Table 8). This is needed for PDT functors but there is no counterpart in AnCora and PropBank.

6.9. Miscellaneous: MISCLL

The **MISCLL** label is designed for miscellaneous relations such as rhematizers, linking and modal adverbial expressions (Table 9).

7. Conclusion

We have surveyed the label inventories of deep-syntactic relations from three annotation frameworks: Meaning-Text Theory, Functional Generative Description and PropBank. Based on these

Label	Meaning	P(CE)DT	AnCora	PropBank
BINDER:conj	conjunction	CONJ		-
BINDER:apps	apposition	APPS		-
BINDER:disj	disjunction	DISJ		-
BINDER:advs	adversative	ADVS		-
BINDER:confr	confrontation	CONFR	N/A	-
BINDER:contra	conflict	CONTRA		-
BINDER:grad	gradation	GRAD		-
BINDER:reas	reason / cause	REAS		-
BINDER:csq	consequence	CSQ		ARGM-CSQ
BINDER:oper	math operation	OPER		-

Table 7: BINDER Role

Label	Meaning	P(CE)DT	AnCora	PropBank
ADNOM:app	appurtenance	APP	-	-
ADNOM:auth	author	AUTH	-	-
ADNOM:id	identity	ID	-	-
ADNOM:mat	content	MAT	-	-
ADNOM:restr	modification	RSTR	-	-

Table 8: ADNOM Role

Label	Meaning	P(CE)DT	AnCora	PropBank
MISCLL:att	speaker's attitude	ATT	-	-
MISCLL:intf	expletive subject	INTF	-	-
MISCLL:mod	modal expression	MOD	-	ARGM-MOD
MISCLL:prec	preceding context	PREC	-	ARGM-DIS
MISCLL:rhe	rhematizer	RHEM	-	ARGM-NEG
MISCLL:cphr	complex predicate	CPHR	-	-
MISCLL:dphr	dependent in idiom	DPHR	-	-
MISCLL:fphr	foreign expression	FPHR	-	-
MISCLL:cm	conjunction modifier	CM	-	-

Table 9: MISCLL Role

observations, we propose a unified relation inventory, which contains unified labels for relations that are similar or equivalent in the three frameworks, and additional labels for relations that are unique, so that annotations can be mapped with minimal information loss. The unified inventory is hierarchical so that less-specific relation types can be mapped and the missing finer distinctions do not have to be guessed.

As a result, the five languages discussed in the paper can be queried using the unified set of semantic role labels. Our future plans involve the application of this label set to all UD treebanks. We intend to combine cross-lingual projection (as in the Universal PropBank project) with heuristics that will use the surface syntax as input (and with valency frame lexicons if available). In any case, such annotation is destined to contain noise. But with the unified label set, we can at least provide comparable annotation for datasets where it has to be estimated automatically, and for those where it can be obtained from dedicated deep syntactic/semantic resources.

Our semantic labels are applicable to enhanced as well as basic relations, and we intend to apply them to enhanced graphs in future work. However, the current dataset is based on basic trees, which are available for all UD languages.

The three current source frameworks (and in particular FGD) have quite detailed inventories of relations, therefore we believe that the proposed universal set already covers a substantial part of what can be found in deep-syntactic datasets in general.

Acknowledgements

This work was supported by the Grant 20-16819X (LUSyD) of the Czech Science Foundation (GAČR); and LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic.

8. Bibliographical References

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

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Kilian Evang. 2023. Superframes for consistent and comprehensive semantic role annotation. In *UniDive 1st General Meeting – Selected Abstracts*, Orsay, France. Université Paris-Saclay. https://unidive.lisn.upsaclay.fr/doku.php?id=meetings:2023-saclay:abstracts.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, M Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. European Language Resources Association.

Katri Haverinen, Jenna Kanerva, Samuel Kohonen, Anna Missilä, Stina Ojala, Timo Viljanen, Veronika Laippala, and Filip Ginter. 2015. The Finnish proposition bank. *Language Resources and Evaluation*, 49:907–926.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Linh Ha, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal Proposition Bank 2.0. In *Proceedings of the 13th Conference on Language Resources*

and Evaluation (LREC 2022), pages 1700–1711, Marseille, France. European Language Resources Association (ELRA).

Sylvain Kahane. 2003. The Meaning-Text Theory. In *To appear in Dependency and Valency. An International Handbook of Contemporary Research*, Berlin, Germany. De Gruyter.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Jarmila Panevová. 1974. On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics*, 22(3-40):6–3.

Petr Sgall. 1967. Functional sentence perspective in a generative description. *Prague studies in mathematical linguistics*, 2(1967):203–225.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, pages 197–209. Springer.

Mariona Taulé, M Antònia Martí, and Oriol Borrega. 2011. AnCora 2.0: Argument structure guidelines for Catalan and Spanish. Technical report, CLiC (Text-Mess 2.0 Project), Barcelona, Spain.

Mariona Taulé, M Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

9. Language Resource References

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, et al. 2013. Prague Dependency Treebank 3.0.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. Prague dependency treebank. *Handbook of Linguistic Annotation*, pages 555–594.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Prague Czech-English Dependency Treebank 2.0.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štepánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševcíková-Razímová, et al. 2006. Prague dependency treebank 2.0.