HamleDT 2.0: Thirty Dependency Treebanks Stanfordized

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, CZ-11800, Czechia
{rosa,masek,marecek,popel,zeman,zabokrtsky}@ufal.mff.cuni.cz

Abstract

We present HamleDT 2.0 (HArmonized Multi-LanguagE Dependency Treebank). HamleDT 2.0 is a collection of 30 existing treebanks harmonized into a common annotation style, the Prague Dependencies, and further transformed into Stanford Dependencies, a treebank annotation style that became popular in recent years. We use the newest basic Universal Stanford Dependencies, without added language-specific subtypes. We describe both of the annotation styles, including adjustments that were necessary to make, and provide details about the conversion process. We also discuss the differences between the two styles, evaluating their advantages and disadvantages, and note the effects of the differences on the conversion. We regard the stanfordization as generally successful, although we admit several shortcomings, especially in the distinction between direct and indirect objects, that have to be addressed in future. We release part of HamleDT 2.0 freely; we are not allowed to redistribute the whole dataset, but we do provide the conversion pipeline.

Keywords: treebanks, Stanford dependencies, harmonization

1. Introduction

Dependency treebanks are available for several dozen languages, and there is a growing interest in multilingual/cross-lingual syntactic parsing experiments. However, the individual treebanks use different annotation schemes. The differences can be found in the dependency-relation label sets (the names and definitions of the labels and the granularity of the set) as well as in the dependency structures capturing coordination, subordinate clauses, verb groups, prepositional phrases and other linguistic phenomena (Zeman et al., 2012). These divergences present a significant obstacle to the use of these resources in multilingual language technologies or for evaluation of cross-lingual syntactic parsers (McDonald et al., 2011). Therefore, a logical step to take is to convert the various treebanks into the same schema.

So far, the largest collection of harmonized treebanks has been HamleDT 1.0, a compilation of 29 existing dependency treebanks (or dependency conversions of other treebanks). The treebanks were harmonized into the Prague Dependencies (PRG) style of annotation (Zeman et al., 2012), which is a slight adaptation of the annotation style of the Prague Dependency Treebank (PDT) (Hajič et al., 2006). We list the source treebanks and describe PRG in Section 2.

Recently, the Stanford Dependencies (SD) representation (de Marneffe et al., 2006; de Marneffe et al., 2013) has gained in popularity and, although primarily defined for English, has been successfully used by researchers in various domains and for various languages. Moreover, Universal Stanford Dependencies (USD) (de Marneffe et al., 2014) have just been introduced, which focus on adapting the previous version to capture grammatical relations across languages. Until now, the largest collection of treebanks subscribing to the SD has been the Universal Dependency Treebank of McDonald et al. (2013), currently including 11 languages (7 of which have been annotated manually directly using SD; the others were converted automatically).

In this paper, we present HamleDT 2.0, a collection of 30 treebanks annotated in basic USD. The treebanks, already harmonized to PRG in HamleDT 1.0, were automatically converted from PRG to USD in a language-independent (and also source-treebank-independent) way, thereby creating the largest existing collection of treebanks annotated in USD. The resource is released on our website.[1] We detail the stanfordization in Section 3. and discuss some encountered issues in Section 4.

2. Harmonization

The first step in creating the data resource presented was the collection of the source treebanks, and their automatic rule-based conversion (harmonization) to PRG, including conversion of part of speech tags and other annotated morphological information into Interset representation (Zeman, 2008). This has already been done in HamleDT 1.0 (Zeman et al., 2012) and further improved in HamleDT 1.5.[1] In HamleDT 2.0, a new Slovak treebank was added.

In this section, we list the source treebanks in Table 1 and describe the current version of PRG used in HamleDT. For details about the harmonization itself, please refer to (Zeman et al., 2012).

As the current version of HamleDT contains just one treebank per language, we use the ISO language codes to refer to individual treebanks throughout this paper. Any claims are to be read as claims about the particular treebank, and not necessarily about the language in general.

2.1. Prague Dependencies

There are at least ten treebanks (both from Prague and from other places) that use the PRG label set natively. Eight of

[1] http://ufal.mff.cuni.cz/hamledt
The licenses allow us to distribute only a subset of the whole data set, consisting of 13 treebanks. For the rest of the treebanks, the user has to obtain the source treebank first; we then provide the conversion tools.

Arabic [ar]: Prague Arabic Dependency Treebank 1.0 / CoNLL 2007 (Smrž et al., 2008) http://padt-online.blogspot.com/2007/01/conll-shared-task-2007.html
Basque [eu]: Basque Dependency Treebank, a larger version than the one included in CoNLL 2007, generously provided by IXA Group (Aduriz et al., 2003) http://hdl.handle.net/10230/17098
Bengali [bn], Hindi [hi] and Telugu [te]: Hyderabad Dependency Treebank / ICON 2010 (Husain et al., 2010) http://ltrc.iiit.ac.in/icon/2010/nlptools/
Bulgarian [bg]: BulTreeBank (Simov and Osenova, 2005) http://www.bultreebank.org/indexBTB.html
Catalan [ca] and Spanish [es]: AnCora (Taulé et al., 2008) http://clic.ub.edu/corpus/en/ancora-descarregues
Czech [cs]: Prague Dependency Treebank 3.0 (Bejček et al., 2013) http://ufal.mff.cuni.cz/pdt3.0/, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3
Danish [da]: Danish Dependency Treebank / CoNLL 2006 (Kromann et al., 2004), now part of the Copenhagen Dependency Treebank http://code.google.com/p/copenhagen-dependency-treebank/
Dutch [nl]: Alpino Treebank / CoNLL 2006 (van der Beek et al., 2002) http://odur.let.rug.nl/~vannoord/trees/
English [en]: Penn TreeBank 3 / CoNLL 2007 (Marcus et al., 1993) http://www.cis.upenn.edu/~treebank/
Estonian [et]: Eesti keele puudepank / Arborest (Bick et al., 2004) http://www.cs.ut.ee/~kaili/Korpus/puud/
Finnish [fi]: Turku Dependency Treebank (Haverinen et al., 2010) http://bionlp.utu.fi/fintreebank.html
German [de]: Tiger Treebank / CoNLL 2009 (Brants et al., 2004) http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
Greek (modern) [el]: Greek Dependency Treebank (Prokopidis et al., 2005) http://gdt.ilsp.gr/
Greek (ancient) [grc] and Latin [la]: Ancient Greek and Latin Dependency Treebanks (Bamman and Crane, 2011) http://nlp.perseus.tufts.edu/syntax/treebank/
Hindi [hi]: *see Bengali*
Hungarian [hu]: Szeged Treebank (Csendes et al., 2005) http://www.inf.u-szeged.hu/projectdirs/hlt/index_en.html
Italian [it]: Italian Syntactic-Semantic Treebank / CoNLL 2007 (Montemagni et al., 2003) http://medialab.di.unipi.it/isst/
Japanese [ja]: Verbmobil (Kawata and Bartels, 2000) http://www.sfs.uni-tuebingen.de/en/tuebajs.shtml
Latin [la]: *see Greek (ancient)*
Persian [fa]: Persian Dependency Treebank (Rasooli et al., 2011) http://dadegan.ir/en/persiandependencytreebank
Portuguese [pt]: Floresta sintá(c)tica (Afonso et al., 2002) http://www.linguateca.pt/floresta/info_floresta_English.html
Romanian [ro]: Romanian Dependency Treebank (Călăcean, 2008) http://www.phobos.ro/roric/texts/xml/
Russian [ru]: Syntagrus (Boguslavsky et al., 2000) http://ruscorpora.ru/en/
Slovak [sk]: Slovak National Corpus (in development) (Šimková and Garabík, 2006) https://metashare.korpus.sk/repository/search/?q=treebank
Slovene [sl]: Slovene Dependency Treebank / CoNLL 2006 (Džeroski et al., 2006) http://nl.ijs.si/sdt/
Spanish [es]: *see Catalan*
Swedish [sv]: Talbanken05 (Nilsson et al., 2005) http://www.msi.vxu.se/users/nivre/research/Talbanken05.html
Tamil [ta]: TamilTB (Ramasamy and Žabokrtský, 2012) http://ufal.mff.cuni.cz/~ramasamy/tamiltb/0.1/
Telugu [te]: *see Bengali*
Turkish [tr]: METU-Sabanci Turkish Treebank (Atalay et al., 2003) http://www.ii.metu.edu.tr/content/treebank

Table 1: List of treebanks included in HamleDT 2.0

them are included in HamleDT 2.0 [cs, ar, el, grc, la, sk, sl, ta]. In addition, this label set is also used in the Prague English Dependency Treebank (PEDT 2.0)[2] and HOBS (Croatian) (Berović et al., 2012).

There are 44 different labels of non-root nodes that appear in at least one of these treebanks. About 15 of them are widely attested in most or all treebanks; the rest can be considered language- or treebank-specific. HamleDT uses 21 labels that occur in [cs], plus the additional labels `Apposition` and `Neg`.

Technically, dependency labels are attributes of child nodes. Most of the time, the label describes the relation between the child and its parent, but there are a few important exceptions: `Pred`, `Coord`, `AuxP`, `AuxC`, `ExD` (see below for details). Furthermore, one attribute is orthogonal to the label space: `is_member` marks members of paratactic structures (conjuncts). Depending on file format, this attribute is stored either separately or as extension of the main labels (e.g. `Pred_M` means that the main label is `Pred` and that it is a conjunct).

The list of PRG labels used in HamleDT follows. Note that clausal and non-clausal dependents are not distinguished.

`Pred` – Main predicate, a node not depending on another node. Note that predicates of subordinate clauses do not get the label `Pred`—their label describes the relation of the clause to its parent.

`Sb` – Subject. Typically a noun phrase in nominative, although infinitives and other realizations are possible.

`Obj` – Object. Besides noun phrases and prepositional phrases, this class also includes infinitives attached to modal verbs.

`Pnom` – Nominal predicate. Typically an adjective, participle or a noun phrase attached to the copula *to be*.

`Atv`, `AtvV` – Determining complement, verbal attribute. A node that depends both on a verb and on its argument (subject or object). As the dependency structure ought to be a tree, the node is technically attached only to one of the parent candidates. Nominal parent (and the `Atv` label) is preferred; if it is not present due to ellipsis, then the node is attached to the verb and labeled `AtvV`. The most prominent examples involve the conjunction *as*: "*as the acting president,* I feel obliged to…" (attached to subject) or "we were preparing for that *as [for] the last chance*" (attached to object).

`Adv` – Adverbial modifier of a verb, adjective, adverb or numeral ("*approximately* ten"). Typically realized as adverb, prepositional phrase or clause.

`Atr` – Attribute of a noun, pronoun or numeral (one number attached to another, e.g. in dates). Realized as adjective, another noun, prepositional phrase or clause. A noun parent should never get an `Adv` child: even "stadium *in London*" will be analyzed as `Atr`.

`Apposition` – A noun phrase attached to another noun phrase as a parenthetical explanation, e.g. "Elisabeth II, *the Queen of England*". This label is intentionally distinct from `Apos`, used in the original treebanks. The attachment of apposition in HamleDT is very similar to nominal attributes, while the original Prague annotation style is to treat it as a

[2] http://ufal.mff.cuni.cz/pedt2.0

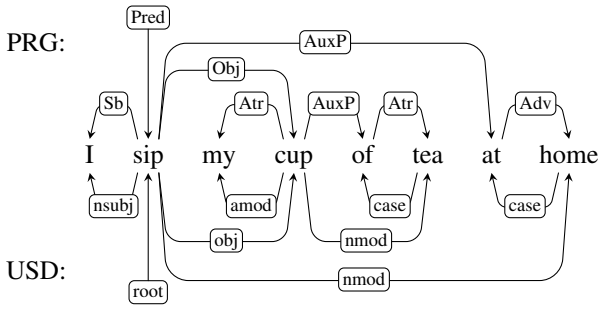

Figure 1: Labeling of prepositional phrases.

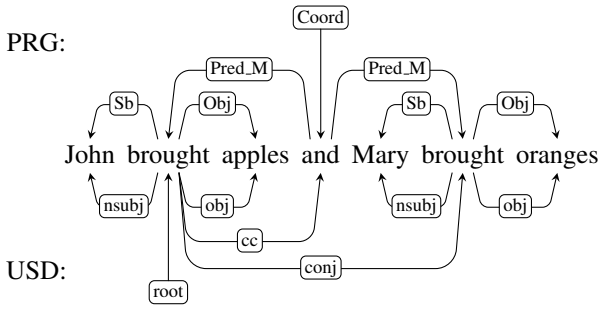

Figure 2: Coordination without ellipsis.

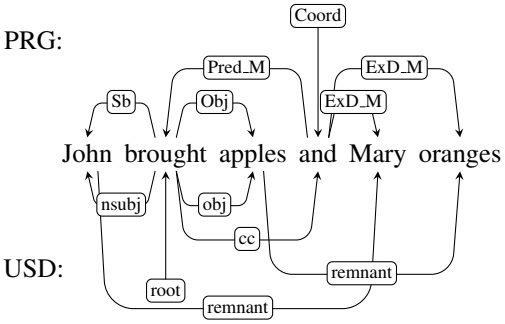

Figure 3: Coordination with ellipsis.

paratactic structure, similar to coordination. This is the first major deviation from the style of PDT. Unlike the creators of PDT, we do not see apposition as an inherently paratactic construction; this is in accord with all the non-Prague treebanks in our collection.

`Coord` – Coordinating node. Usually a conjunction, sometimes comma or other punctuation. This label does not describe the relation of the node to its parent. Such relations are marked at the conjuncts. Children of the `Coord` node are classified as either conjuncts or shared dependents of the conjuncts. (Every conjunct may have its own private dependents in addition to the shared ones.)

`AuxP` – Primary preposition or part of secondary (compound) preposition. Preposition determines the case of the noun phrase in many Indo-European languages, therefore it is annotated as the parent of the noun in PRG. Still, it is considered an auxiliary node and the real lexical dependency goes from the parent of the preposition to its noun child (this is driven by the same lexicalist principle that led in USD to attaching prepositions as child nodes of their nouns). The function of the whole prepositional phrase is annotated at the noun. See Figure 1.

`AuxC` – Subordinating conjunction. Parallel to prepositions, subordinating conjunctions are auxiliary nodes on the path between the predicate of the subordinate clause and its governor. The real function of the clause is annotated at the predicate.

`AuxV` – Auxiliary verb attached to main verb.

`Neg` – Particle that negates the meaning of its parent. Language-specific, as some languages express negation using bound morphemes. This label is not used in PDT (it is not needed for Czech) but it is useful in quite a few languages and we recently introduced it in HamleDT. We still do not recognize it in all the treebanks where it would apply. In some of the treebanks, negative particles are labeled as adverbial modifiers.

`AuxT` – In PDT, this label is reserved for reflexive pronouns attached to inherently reflexive verbs *(reflexiva tantum)*. In HamleDT, we also use it for particles that modify the meaning of verb, e.g. "make *up*".

`AuxR` – Reflexive pronoun used to form reflexive passive, as in [cs] "to *se* udělá snadno" ([en] "it is done easily"; lit. "it *itself* does easily").

`AuxO` – Redundant or emotional item, redundant coreferential pronoun.

`AuxZ` – Emphasizing word ("*especially* on Monday").

`AuxX` – Comma that does not serve as head of coordination.

`AuxK` – Sentence-final punctuation such as period, exclamation or question mark. Not quotation mark or bracket. Periods that serve both as abbreviation markers and sentence terminators are analyzed as abbreviation, i.e. they are attached to the abbreviated word and labeled `AuxG`.

`AuxG` – Graphic symbol other than comma or sentence-final punctuation.

`AuxY` – Garbage can for adverbs, conjunctions, particles etc. that do not qualify for a better label. Example: *actually*. This label is also used for dependent parts of multiword conjunctions (such as *such as*), extra conjunctions in coordinations etc.

`ExD` – Externally dependent (see Figures 2 and 3). This label in general signals that the structural information is incomplete. If an item was elided but its dependents remain in the sentence, they get `ExD`. Example: *John brought some apples and Mary* [brought] *some oranges. Mary* and *oranges* cannot be attached to the second *brought* because it is missing. They are therefore treated as conjuncts, together with the first *brought*. They cannot be labeled as predicates (which they are not), nor can they get `Sb`/`Obj` (the visible dependency link does not lead to the verb of which they are subject/object). So they get the technical label `ExD`.

3. Stanfordization

3.1. Universal Stanford Dependencies in HamleDT

In the structure of the dependency tree, we try to fully adhere to USD as defined by de Marneffe et al. (2014). In copular constructions, the nominal predicate is the head governing the copula as a complement. The adpositions (and any other case-marking functional words) and subordinating conjunctions are structurally treated the same; they are governed by the word they introduce (see Figure 1). In

coordination constructions, the conjuncts are siblings except for the first one, which is the head, and the shared modifiers and coordinating conjunctions are governed by the head conjunct – see Figure 2 for an example.

We base our label set on USD; however, we add one new label, and a few other labels remain unused. The reason is not differing linguistic views, but rather a technical one – the relations marked by the unused labels cannot be reliably distinguished from others in PRG in a language-independent way, and some of them are indistinguishable even in the source treebanks. Figure 4 shows the hierarchy of USD with our modifications; additional labels are marked by $^{+}$ and the labels not used by us are marked by $^{\times}$. The reasoning behind these differences is given in Section 4.

USD assume the extension of the label set with language-particular labels as subtypes of existing labels. We leave this for future work; in HamleDT 2.0, we are only trying to convert the treebanks from PRG to the general label set, which should be a basis for future extensions thereof.

3.2. The conversion

Stanfordization of the harmonized treebanks consists of rehanging some of the nodes, and mapping the PRG labels to USD labels. The pipeline is implemented in the Treex framework[3] and is language-independent, as all the language-specific and treebank-specific conversions take place already in the harmonization phase (Section 2.).

Structural changes were performed for constructions that involve adpositions, copula verbs, coordinations, subordinations, and punctuation – in all of these cases, USD differ from PRG in their principle of nodes considered as auxiliary being leaf nodes.

A USD label for a node is devised based mainly on its PRG label and its part of speech, often also taking into account the label and part of speech of its parent, and, with verbs, the finiteness and passiveness marked in the Interset morphological representation. A basic overview of the mapping is provided in Table 2.

4. Discussion

In this section, we discuss the differences between USD and PRG, the effect of these differences on the process of stanfordization of HamleDT, and specify resulting implications for future work.

Both PRG and USD have a common goal of providing a linguistic formalism that can capture the syntax of sentences across diverse languages. Both of them were originally designed primarily for one language only (SD for English, PRG for Czech) and had to be adapted for the multilingual setting; in both cases most of the adaptation happens in the label set, while the tree structure remains largely unchanged.

SD were adapted to USD by simplifying and generalizing the original label set to capture important phenomena attested in a significant proportion of languages, intricately building upon the hierarchical structure of the labels to allow and encourage language-specific extensions of the label set by adding subtypes of the universal types, e.g. `poss` for possessiveness marking ("'s" in English) as a subtype of `case`. This allows great flexibility for accurately representing the sentences of various languages while always

Core dependents of clausal predicates
- `csubj` – clausal subject
  - `csubjpass` – passive clausal subject
- `nsubj` – non-clausal subject
  - `nsubjpass` – passive non-clausal subject
- $^{+}$`obj` – object
  - $^{\times}$`dobj` – direct object
  - $^{\times}$`iobj` – indirect object
- `ccomp` – clausal complement
- `xcomp` – open clausal complement

Non-core dependents of clausal predicates
- `nmod` – nominal modifier
- `advcl` – adverbial clause modifier
- `nfincl` – non-finite clause modifier
- `advmod` – adverbial modifier
- `neg` – negation modifier
- $^{\times}$`ncmod` – nominalized clause modifier

Special clausal dependents
- `aux` – auxiliary
  - `auxpass` – passive auxiliary
  - `cop` – copula
- `mark` – marker introducing `advcl` or `ccomp`
- `punct` – punctuation
- $^{\times}$`vocative` – vocative
- $^{\times}$`discourse` – discourse elements
- $^{\times}$`expl` – expletives and frozen reflexive

Coordination
- `conj` – non-first conjunct
- `cc` – coordinating conjunction

Noun dependents
- `nummod` – numeric modifier
- `appos` – appositional modifier
- `nmod` – nominal modifier
- `relcl` – relative clause modifier
- `nfincl` – non-finite clause modifier
- `amod` – adjectival modifier
- `det` – determiner
- `neg` – negation modifier
- $^{\times}$`ncmod` – nominalized clause modifier

Compounding and unanalyzed
- `mwe` – part of fixed grammaticized expression
- `compound` – other multi-word lexemes
- $^{\times}$`name` – multiple nominal word proper nouns
- $^{\times}$`foreign` – sequence of foreign words
- $^{\times}$`goeswith` – part of the governing word

Case-marking, prepositions, possessive
- `case` – case marker (e.g. adpositions)

Loose joining relations
- `remnant` – child of an elided node
- $^{\times}$`parataxis` – parataxis
- $^{\times}$`list` – list member
- $^{\times}$`reparandum` – speech repair
- $^{\times}$`dislocated` – preposed/postposed element

Other
- `root` – sentence head
- `dep` – general dependent

Figure 4: USD as used in HamleDT 2.0.

[3] `http://ufal.mff.cuni.cz/treex/`

PRG	USD
Pred	root
Sb	nsubj, csubj, nsubjpass, csubjpass
Obj, Pnom, Atv, AtvV, AuxR	obj, ccomp, xcomp
Adv, AuxO, AuxY, AuxZ	advmod, nmod, advcl, nfincl, mwe
Atr	amod, nmod, nummod, relcl, nfincl
AuxA	det
Neg	neg
AuxV	aux, auxpass, cop
AuxP	case, mwe
AuxC	mark
Apposition	appos
AuxT	mwe
ExD	remnant
Coord	cc
non-first conjunct (any label)	conj
punctuation (any label)	punct

Table 2: A basic overview of mapping PRG labels to USD labels.

having a common backbone of the universal types. (We have not yet resorted to introducing such subtypes in stanfordized HamleDT, as we decided to first focus on having at least the common backbone.)

The adaptation of PRG labels in HamleDT has been considerably less systematic, trying to map the phenomena in the various languages onto the PRG labels by choosing one closest in its definition, introducing a new label only very rarely. This is one of the reasons that harmonization to PRG labels leads to loss of information in some cases, such as in case of the direct/indirect object distinction which PRG does not capture even if the source treebank does.

The two formalisms themselves differ in many aspects, especially because of the following reasons:

The underlying theory. PRG build upon the Function-Generative description by Sgall (1967), which understands the shallow syntax tree as only one of the annotation layers, built on morphological layer and leading to the tectogrammatical (deep-syntax) layer. USD build upon Lexical-Functional Grammar of Bresnan (2001).

The main goal. USD try to make the representation easy to use for the user in applications such as Information Retrieval, while PRG aim at being as linguistically accurate as possible, even providing quite complex annotation rules rather than resorting to simplifications.

Thus, converting PRG annotations to USD annotations is not a trivial task. While there are some cases of 1:1 mapping of labels with no rehanging of nodes, this is usually not the case, and one has to be creative when trying to map the labels; the transformations of tree structure are less common and usually easier to perform. We therefore broadened the definitions of some of the labels to allow us to use them for many less-common phenomena, and we currently have one extra label and 13 unused labels from the perspective of USD. We illustrate some non-trivial conversion issues in the following paragraphs; a graphical representation of frequencies of the individual labels in the treebanks after stanfordization is shown in Figure 5.

In future, we plan to investigate whether the source treebanks contain information that is lost in harmonization and is then missing for accurate stanfordization, which will probably lead to enrichment of PRG so that no information deemed important is lost and the stanfordization can remain language-independent. In this paper, however, we limit ourselves to exploring the extent to which USD representation can be devised from the current PRG.

4.1. Coordinations

Coordinating structures are known to be difficult to capture by dependency trees, as the coordination itself is not a dependency relation. The properties of coordination representation in PRG and SD and conversions thereof have already been studied by Popel et al. (2013), showing PRG to have more expressive power than SD; thus, conversion of coordinating structures from PRG to USD is easy, although lossy.

The main advantage of PRG over (basic) SD is the ability to distinguish private and shared modifiers. It is arguable whether this distinction should be captured in syntactic annotation, because the construction is often truly ambiguous, as in "*green* tables and chairs". In other cases, disambiguation can be done based on semantics, as in "juice from *black* currant and bananas" versus "we serve *fish* steaks and fingers"; it is still questionable whether the disambiguation is to be made on syntactic level, but the same can be said about e.g. PP attachment. However, there are still other cases, where the construction is clearly non-ambiguous, as in "*Peter* plays and sings" vs "*Peter* plays and Mary sings". Because of this, we believe that an annotation style that can capture the private/shared modifier distinction is superior to one that can not.

On the other hand, as many of the source treebanks do not use PRG, the aforementioned distinction is often not present in the source treebank in the first place, and heuristics had to be used to convert these into PRG. From that point of view, SD are a better fit, as no heuristics are necessary to convert most of the coordinations into them.

In future, we would like to explore the extended SD as well (de Marneffe and Manning, 2008), as these are able to make the private/shared modifier distinction, among other benefits they bring.

4.2. Objects (adding a label)

USD distinguish direct and indirect objects, which PRG do not. De Marneffe et al. (2014) make a point by attesting that this distinction can be traced in many languages, while in PRG, such distinctions are not made, as the notion of direct and indirect object is not usual in Czech linguistics. We want to address this shortcoming in future by exploring the source treebanks and extending PRG appropriately to capture object type distinctions that are common across languages. Currently, we simply introduce an additional `obj` label, which we use instead of `dobj` and `iobj`.

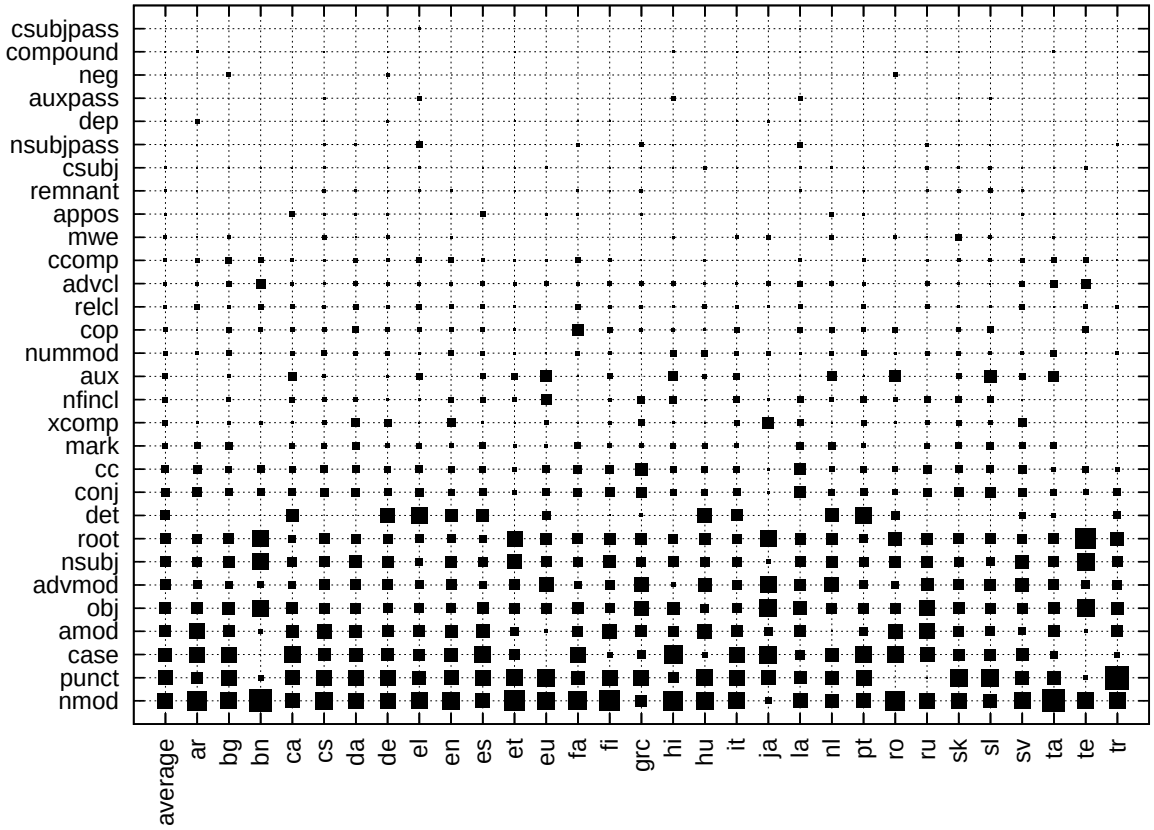

Figure 5: Distribution of USD labels in all the 30 HamleDT treebanks after stanfordization. The area of each square corresponds to the proportion of the given label in the given language.

4.3. Currently unused labels

There are thirteen USD labels we do not use. They can be divided into the following categories:

- `dobj`, `iobj`, `name`, `foreign`, and `vocative` are unused because we cannot make the necessary distinctions based on the current PRG harmonization. We are planning to revisit the harmonization process to see whether we could get the necessary information from the source treebanks, as we find these labels useful.

- `discourse`, `goeswith`, `reparandum`, and `list` also cannot be distinguished in PRG; however, we are still considering whether to try to include these labels in future or not, as we find them to be of little importance. We believe that using such labels is useful when annotating a new treebank, as the respective phenomena are typically hard to label consistently, but if the source treebank did not choose to use them, it will be nearly impossible to detect and label them in the harmonization; moreover, the added value brought by such an effort would be rather limited in our opinion.

- `ncmod`, `expl`, `dislocated`, and `parataxis` are currently unused, as we are unsure about their exact definition in the multilingual setting. We will reconsider their inclusion once this is clearer.

4.4. Broadening the definitions of labels

During the conversion, we found that the definitions of labels by de Marneffe et al. (2014) are rather narrow, for instance frequently distinguishing clausal and nominal elements but disregarding other possibilities. This is insufficient for many languages, including English – consider e.g. the sentence "Quickly does not always mean efficiently.", where both the subject and object are adverbs and should probably be classified neither as nominal nor clausal. We therefore broaden the definition of `nsubj` and `[di]obj` from "nominal" to "non-clausal" subject and object, as we believe the clausal/non-clausal distinction to be of a higher importance than the nominal/non-nominal distinction. Similar generalizations are made for other labels where necessary, which we prefer to defining new labels that would be only slightly different from the existing ones and at the same time very rare in the data. In future, such distinctions may be made by defining new subtypes of existing labels if the current label set is found to be too coarse.

4.5. Clauses

It is a non-trivial task to correctly identify and classify clauses. Following the definitions of de Marneffe et al. (2014), we treat every non-auxiliary verb as a clause head. PRG do not help us in classifying a clause as finite or non-finite; therefore, we resort to inspecting the Interset morphological representation of the clause head, and annotate a clause as finite if it is headed by a verb form marked as finite. This is clearly an approximation, as even finite clauses can be headed by infinite verb forms in many languages, their finiteness being marked by the auxiliary verbs; in other languages, the infiniteness is marked by a particle, similar to English "to". Consequently, the morphological annotation in some source treebanks does not always dis-

tinguish finite and infinite verb forms; currently, we treat the verb as finite in such cases, as we have no language-independent way of making the distinction. It is also not clear how to treat verb forms that are marked as neither finite nor infinite, such as participles or transgressives; following de Marneffe et al. (2014), we currently treat them as infinite, but this might not be appropriate for all languages. Thus, the necessary distinction is currently hard to make in a language-independent way, and it is clear that the so far rather neglected harmonization of verbs in HamleDT will have to be improved.

5. Conclusion

In this paper, we presented HamleDT 2.0, a collection of 30 pre-existing treebanks, which we automatically harmonized into PRG, and subsequently converted to USD. We encountered several issues during the conversion process, leading us to make slight modifications to the USD definition and leave some future work for us in improving both the harmonization step and the conversion step.

We release 13 of the harmonized and converted treebanks at `http://ufal.mff.cuni.cz/hamledt`. The licenses for the rest of the original treebanks do not allow us to redistribute them, but our harmonization and conversion pipeline is available freely.

6. Acknowledgements

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap), SVV 260 104, Czech Science Foundation grant no. P406/14/06548P, and GAUK 1572314. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

We would also like to thank de Marneffe et al. (2014) for providing us with a draft of their paper, and apologize for any inaccuracies in referring to it due to its possible later changes that we were not aware of.

7. References

Aduriz, Itzair, Aranzabe, María Jesús, Arriola, Jose Mari, Atutxa, Aitziber, Díaz de Ilarraza, Arantza, Garmendia, Aitzpea, and Oronoz, Maite. (2003). Construction of a Basque dependency treebank. In *Proc. of the 2nd Workshop on Treebanks and Linguistic Theories*.

Afonso, Susana, Bick, Eckhard, Haber, Renato, and Santos, Diana. (2002). "Floresta sintá(c)tica": a treebank for Portuguese. In *Proc. of LREC*, pages 1968–1703.

Atalay, Nart B., Oflazer, Kemal, Say, Bilge, and Inst, Informatics. (2003). The annotation process in the Turkish treebank. In *Proc. of the 4th Intern. Workshop on Linguistically Interpreteted Corpora (LINC)*.

Bamman, David and Crane, Gregory. (2011). The Ancient Greek and Latin dependency treebanks. In Sporleder, Caroline, Bosch, Antal, and Zervanou, Kalliopi, editors, *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98. Springer Berlin Heidelberg.

Bejček, Eduard, Hajičová, Eva, Hajič, Jan, Jínová, Pavlína, Kettnerová, Václava, Kolářová, Veronika, Mikulová, Marie, Mírovský, Jiří, Nedoluzhko, Anna, Panevová, Jarmila, Poláková, Lucie, Ševčíková, Magda, Štěpánek, Jan, and Zikánová, Šárka. (2013). Prague dependency treebank 3.0.

Berović, Daša, Željko Agić, and Tadić, Marko. (2012). Croatian dependency treebank: Recent development and initial experiments. In *Proc. of LREC'12*, İstanbul, Turkey. European Language Resources Association (ELRA).

Bick, Eckhard, Uibo, Heli, and Müürisep, Kaili. (2004). Arborest – a VISL-style treebank derived from an Estonian constraint grammar corpus. In *Proc. of Treebanks and Linguistic Theories*.

Boguslavsky, Igor, Grigorieva, Svetlana, Grigoriev, Nikolai, Kreidlin, Leonid, and Frid, Nadezhda. (2000). Dependency treebank for Russian: Concept, tools, types of information. In *Proc. of the 18th conference on Computational linguistics-Volume 2*, pages 987–991. Association for Computational Linguistics Morristown, NJ, USA.

Brants, Sabine, Dipper, Stefanie, Eisenberg, Peter, Hansen, Silvia, König, Esther, Lezius, Wolfgang, Rohrer, Christian, Smith, George, and Uszkoreit, Hans. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, 2(4):597–620. Special Issue.

Bresnan, Joan. (2001). *Lexical-functional syntax*. Wiley-Blackwell.

Csendes, Dóra, Csirik, János, Gyimóthy, Tibor, and Kocsor, András. (2005). The Szeged treebank. In Matoušek, Václav, Mautner, Pavel, and Pavelka, Tomáš, editors, *TSD*, volume 3658 of *Lecture Notes in Computer Science*, pages 123–131. Springer.

Călăcean, Mihaela. (2008). Data-driven dependency parsing for Romanian. Master's thesis, Uppsala University.

de Marneffe, Marie-Catherine and Manning, Christopher D. (2008). The Stanford typed dependencies representation. In *Coling 2008: Proc. of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. ACL.

de Marneffe, Marie-Catherine, MacCartney, Bill, and Manning, Christopher D. (2006). Generating typed dependency parses from phrase structure parses. In *Proc. of LREC*, volume 6, pages 449–454.

de Marneffe, Marie-Catherine, Connor, Miriam, Silveira, Natalia, Bowman, Samuel R., Dozat, Timothy, and Manning, Christopher D. (2013). More constructions, more genres: Extending Stanford dependencies. In *Proc. of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 187–196. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

de Marneffe, Marie-Catherine, Silveira, Natalia, Dozat, Timothy, Haverinen, Katri, Ginter, Filip, Nivre, Joakim, and Manning, Christopher D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proc. of LREC'14*, Reykjavík, Iceland. European Language

Resources Association (ELRA).

Džeroski, Sašo, Erjavec, Tomaž, Ledinek, Nina, Pajas, Petr, Žabokrtský, Zdeněk, and Žele, Andreja. (2006). Towards a Slovene dependency treebank. In *Proc. of LREC 2006*, pages 1388–1391, Genova, Italy. European Language Resources Association (ELRA).

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

Haverinen, Katri, Viljanen, Timo, Laippala, Veronika, Kohonen, Samuel, Ginter, Filip, and Salakoski, Tapio. (2010). Treebanking Finnish. In Dickinson, Markus, Müürisep, Kaili, and Passarotti, Marco, editors, *Proc. of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 79–90.

Husain, Samar, Mannem, Prashanth, Ambati, Bharat, and Gadde, Phani. (2010). The ICON-2010 tools contest on Indian language dependency parsing. In *Proc. of ICON-2010 Tools Contest on Indian Language Dependency Parsing*, Kharagpur, India.

Kawata, Yasuhiro and Bartels, Julia. (2000). Stylebook for the Japanese treebank in Verbmobil. In *Report 240*, Tübingen, Germany.

Kromann, Matthias T., Mikkelsen, Line, and Lynge, Stine Kern. (2004). Danish dependency treebank.

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

McDonald, Ryan, Petrov, Slav, and Hall, Keith. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. ACL.

McDonald, Ryan, Nivre, Joakim, Quirmbach-Brundage, Yvonne, Goldberg, Yoav, Das, Dipanjan, Ganchev, Kuzman, Hall, Keith, Petrov, Slav, Zhang, Hao, Täckström, Oscar, et al. (2013). Universal dependency annotation for multilingual parsing. In *Proc. of ACL*.

Montemagni, Simonetta, Barsotti, Francesco, Battista, Marco, Calzolari, Nicoletta, Corazzari, Ornella, Lenci, Alessandro, Zampolli, Antonio, Fanciulli, Francesca, Massetani, Maria, Raffaelli, Remo, Basili, Roberto, Pazienza, Maria Teresa, Saracino, Dario, Zanzotto, Fabio, Mana, Nadia, Pianesi, Fabio, and Delmonte, Rodolfo. (2003). Building the Italian syntactic-semantic treebank. In Abeillé, Anne, editor, *Building and using Parsed Corpora*, Language and Speech series, pages 189–210, Dordrecht. Kluwer.

Nilsson, Jens, Hall, Johan, and Nivre, Joakim. (2005). MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. In *Proc. of the NODALIDA Special Session on Treebanks*.

Popel, Martin, Mareček, David, Štěpánek, Jan, Zeman, Daniel, and Žabokrtský, Zdeněk. (2013). Coordination structures in dependency treebanks. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 517–527, Sofia, Bulgaria. ACL.

Prokopidis, Prokopis, Desipri, Elina, Koutsombogera, Maria, Papageorgiou, Harris, and Piperidis, Stelios. (2005). Theoretical and practical issues in the construction of a Greek dependency treebank. In *Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 149–160.

Ramasamy, Loganathan and Žabokrtský, Zdeněk. (2012). Prague dependency style treebank for Tamil. In *Proc. of LREC 2012*, İstanbul, Turkey.

Rasooli, Mohammad Sadegh, Moloodi, Amirsaeid, Kouhestani, Manouchehr, and Minaei-Bidgoli, Behrouz. (2011). A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank. In *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 227–231, Poznań, Poland.

Sgall, Petr. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

Simov, Kiril and Osenova, Petya. (2005). Extending the annotation of BulTreeBank: Phase 2. In *The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 173–184, Barcelona.

Smrž, Otakar, Bielický, Viktor, Kouřilová, Iveta, Kráčmar, Jakub, Hajič, Jan, and Zemánek, Petr. (2008). Prague Arabic dependency treebank: A word on the million words. In *Proc. of the Workshop on Arabic and Local Languages (LREC 2008)*, pages 16–23, Marrakech, Morocco. European Language Resources Association.

Šimková, Mária and Garabík, Radovan. (2006). Sintaksičeskaja razmetka v slovackom nacional'nom korpuse (Синтаксическая разметка в Словацком национальном корпусе). In *Trudy meždunarodnoj konferencii Korpusnaja lingvistika (*Труды международной конференции Корпусная лингвистика*) – 2006*, pages 389–394, Sankt-Peterburg, Russia. St. Petersburg University Press.

Taulé, Mariona, Martí, Maria Antònia, and Recasens, Marta. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *LREC*. European Language Resources Association.

van der Beek, Leonoor, Bouma, Gosse, Daciuk, Jan, Gaustad, Tanja, Malouf, Robert, van Noord, Gertjan, Prins, Robbert, and Villada, Begoña. (2002). Chapter 5. the Alpino dependency treebank. In *Algorithms for Linguistic Processing NWO PIONIER Progress Report*, Groningen, The Netherlands.

Zeman, Daniel, Mareček, David, Popel, Martin, Ramasamy, Loganathan, Štěpánek, Jan, Žabokrtský, Zdeněk, and Hajič, Jan. (2012). HamleDT: To parse or not to parse? In *Proc. of LREC'12*, İstanbul, Turkey. European Language Resources Association (ELRA).

Zeman, Daniel. (2008). Reusable tagset conversion using tagset drivers. In *Proc. of LREC 2008*, pages 28–30, Marrakech, Morocco. European Language Resources Association (ELRA).