The Prague Bulletin of Mathematical Linguistics
NUMBER 120 APRIL 2023

EDITORIAL BOARD

Published twice a year by Charles University (Prague, Czech Republic)

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 120 APRIL 2023

CONTENTS

Articles

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 120 APRIL 2023 5–30

Prague to Penn Discourse Transformation

Jiří Mírovský, Magdaléna Rysová, Pavlína Synková, Lucie Poláková

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Abstract

The Prague and Penn styles of discourse annotation are close to each other in basic theoretical views and also in taxonomies of semantic types of discourse relations. A transformation from one of the annotation styles to the other should seemingly be a straightforward process. And yet, slight differences in the taxonomies and significant differences in the technical approaches present several interesting theoretical and practical challenges. The paper focuses on handling the most important issues in the transformation process from the Prague style to the Penn style of discourse annotation, in an effort to bring a valuable data resource – the Prague Discourse Treebank – closer to the international scientific community.

1. Introduction

Manually annotated text corpora have proven to be a multilateral and valuable resource for theoretical linguistic research, as well as for applied natural language processing (NLP), both as test data and for training machine-learning algorithms. The usefulness in the latter, however, has been multiplied in recent years with emergence of pre-trained deep learning methods and tools that use large unannotated data – raw texts – for training word embeddings (representation of (sub)words in a high-dimensional vector space) and for pre-training a deep neural network to "understand" basic language properties. Such a pre-trained system allows to fine-tune the model for a highly specific NLP task using a relatively small manually annotated data, leading to state-of-the-art results in many areas of NLP, as was first demonstrated with

 Corresponding author: mirovsky@ufal.mff.cuni.cz
Cite as: Jiří Mírovský, Magdaléna Rysová, Pavlína Synková, Lucie Poláková. Prague to Penn Discourse Transformation. The Prague Bulletin of Mathematical Linguistics No. 120, 2023, pp. 5–30.
doi: 10.14712/00326585.025.

system BERT by Devlin et al. (2019).[1] Similar approach has since been successfully used for many other NLP tasks, including tasks closely related to text coherence, and specifically discourse relations.

The term *discourse relations* refers to semantic relations that connect two discourse units – segments of text expressing mostly individual events, states, situations (Zikánová et al., 2015). In Example 1, a discourse relation holds between two clauses and is signalled by an explicit discourse-structuring device, a connective *but*.

(1) *Profit may be low*, but **at least costs should be covered**. (PDTB, wsj_0051)
[*Zisk může být malý*, ale **měly by se alespoň zaplatit náklady**.[2] (PCEDT, wsj_0051)]

Depending on a chosen taxonomy, a discourse relation can be classified in one of (usually several tens of) semantic types (e.g., in Example 1, *Comparison.Concession.Arg2-as-denier*, or in another taxonomy, *opposition*). If a discourse relation is marked by a connective, we call it an *explicit* discourse relation. If the connective is absent, we call the relation *implicit*.

A growing interest in text coherence-aware methods can be traced in many areas of natural language processing, including tasks such as machine translation (Xiong et al., 2019; Meyer and Webber, 2013), text generation (Kiddon et al., 2016), summarization (Zhang, 2011), information extraction, opinion mining (Turney and Littman, 2003), coherence evaluation (Rysová et al., 2016), or machine translation evaluation (Bojar et al., 2018). Many of these tasks incorporate a discourse parser in the text pre-processing and, of course, discourse parsing methods have received a lot of attention from the NLP community, including two CoNLL shared tasks (Xue et al., 2015, 2016). Recently, pre-trained deep learning systems such as BERT have spread also to this field: Shi and Demberg (2019) use BERT for classification of so-called implicit discourse relations, outperforming the state of the art. Similarly, Mírovský and Poláková (2021) show that information about the presence of a discourse connective can be incorporated into the BERT framework and that text corpora annotated manually with explicit discourse relations can be successfully used to fine-tune BERT to classify also explicit discourse relations (both in Czech and English).

Several theoretical frameworks for discourse relations representation were developed and used both for theoretical description and for corpora annotation in last decades, with two of them being probably most influential: the approach developed and first used for the annotation of the Penn Discourse Treebank (PDTB; Prasad et al.,

[1] The authors used BERT to reach or improve state-of-the-art results for tasks such as language understanding, question answering and language generation.

[2] We adopt here the Penn Discourse Treebank convention of highlighting two discourse arguments and the connective - Argument 1 (the left one in coordinated structures or in inter-sentential relations, or the governing one in subordinated structures) is typeset in italics, Argument 2 (the other argument) in bold and the connective is underlined.

2008; Prasad et al., 2019), and the Rhetorical Structure Theory (RST; Mann and Thompson, 1988; Taboada and Mann, 2006). While the PDTB model works "locally", i.e. it looks for discourse relations between two (mostly) adjacent clauses or sentences, the RST represents a "global" coherence model, considering each document as a whole to be hierarchically interconnected by rhetorical relations, forming a single tree-like structure.

The Prague Discourse Treebank (PDiT, Poláková et al., 2013; Rysová et al., 2016) is a large corpus of Czech newspaper texts manually annotated with discourse relations. The annotation of discourse relations in PDiT adopts the "local" approach to discourse relations representation and in many aspects is similar to the PDTB approach and is inspired by it (see Section 2). In fact, the relative theory-neutrality of the PDTB approach, the easy applicability of its annotation scheme also to languages other than English, a usually fair inter-annotator agreement and – given its relative simplicity – the possibility to manually annotate a relatively large text corpus, attracted many followers and has been employed in numerous annotation projects.[3] Also both CoNLL shared tasks mentioned above used data annotated according to the PDTB principles.

However, in the Prague Discourse Treebank, unlike most other discourse-annotated corpora, the annotation was not done on top of raw texts but instead on dependency trees of a deep-syntactic layer called *tectogrammatics*. It brings numerous advantages (resolved ellipses, arguments corresponding to subtrees, some relations already captured in the syntax tree, see Mírovský et al. (2012) for details). Yet, a substantial complexity of the native data format of PDiT presents a serious hindrance for any researcher not familiar with the data format and with the annotation theory of the deep-syntactic (tectogrammatical) layer of the corpus.

The present paper deals with theoretical and practical issues of the transformation of the discourse relations annotation of the Prague Discourse Treebank from its original (Prague) format and formalism to the Penn Discourse Treebank framework. In Section 2, we briefly describe the two involved discourse annotation frameworks – the *Penn style* and the *Prague style*. In Section 3, we describe in detail transformation steps from the Prague taxonomy of semantic types (called *discourse types*) to the Penn taxonomy of semantic types (called *senses*). In Section 4, we evaluate the results of the transformation and discuss main differences in sense distributions in the transformed PDiT vs. the PDTB. We conclude and outline future directions in Section 5.

2. Prague and Penn Styles of Discourse Annotation

This section shortly describes relevant parts of the two discourse annotation frameworks under consideration, i.e. the Penn style used in the Penn Discourse Treebank (PDTB), and the Prague style used in the Prague Discourse Treebank (PDiT). We start

[3] Prasad et al. (2008, 2019) (English), Oza et al. (2009) (Hindi), Zeyrek and Kurfalı (2017) (Turkish), Danlos et al. (2012) (French), Zhou and Xue (2012) (Chinese), and many others.

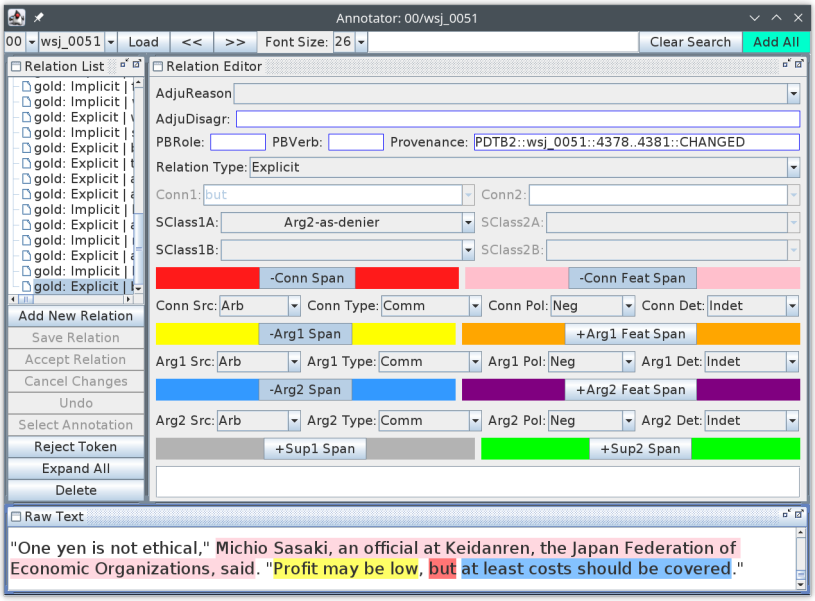


Figure 1. Annotation of the sentence from Example 1 in the PDTB annotation tool Annotator

with the Penn style and follow with the Prague style and its main differences from the Penn style. To easily distinguish the two taxonomies of semantic types in the subsequent text, we use the term *sense* for a semantic type in the Penn style, and the term *discourse type* for a semantic type in the Prague style of discourse annotation.

2.1. Penn Style of Discourse Annotation

The Penn style of discourse annotation employed in the PDTB follows a lexically-grounded approach to annotation of discourse relations (Webber et al., 2003): A discourse connective is a lexical anchor of a discourse relation that holds between two text spans called arguments. The annotation follows the *minimality principle*: the extent of the arguments is marked only as large as needed to interpret the discourse relation properly.

The connective signals the sense of the discourse relation; if it is absent, the relation is called *implicit*. The sense taxonomy is organized into three levels, with four major

classes on the first level and 35 detailed senses[4] on the third level, which also reflects the asymmetry of some of the senses. Table 2 (see Section 4 below) lists all senses for explicit discourse relations in the PDTB 3.0.

If applicable, discourse relations can carry additional information: (i) a *second sense*, if it is distinctly present in the relation beside the first, most prominent sense, (ii) an *attribution* of the relation and of the arguments (i.e., parts of the text that indicate the authors of the statements represented by the relation/arguments), and (iii) a *supplement*, i.e. additional pieces of text beyond the minimality principle that play a supplementary role in interpreting the discourse relation.

In the PDTB 3.0, discourse relations are marked in a stand-off way on top of plain texts (i.e., no text pre-processing needed), and the two arguments, the connective (if present) and other properties are delimited using links to the plain text, i.e. as text spans. In total, there are approx. 25 thousand explicit discourse relations annotated in the PDTB 3.0.

Figure 1 shows the annotation of the discourse relation in the sentence from Example 1 in the PDTB 3.0, displayed in the PDTB annotation tool Annotator (for details on the tool, see Lee et al., 2016).

2.2. Prague Style of Discourse Annotation

Annotation of discourse relations in Czech was to a great extent inspired by the PDTB approach (Poláková et al., 2013). The Prague style of discourse annotation follows the Penn style in marking discourse connectives, their two arguments and the relation semantics, and it also follows the minimality priciple. The list of semantic types of discourse relations (*discourse types*) is close to the list of senses used in the PDTB (especially to the PDTB 3.0 hierarchy), slightly adapted according to the Czech syntactic tradition.[5] The Czech tradition of dependency treebanking was embraced also by incorporating the discourse annotation into the stratificational system of a multilayered language description. Discourse relations thus have not been annotated on plain texts but instead on top of the deep-syntactic (tectogrammatical) layer of the underlying corpus, the Prague Dependency Treebank (PDT; its most recent version was published as a part of the Prague Dependency Treebank - Consolidated 1.0, Hajič et al., 2020).

The underlying corpus, the PDT, is a richly annotated language resource with a multi-layer annotation architecture: (i) a word layer (w-layer), where the plain text is segmented into documents and paragraphs and tokenized, (ii) a morphological layer (m-layer) with segmentation to sentences, all tokens get a lemma and a morphological

[4] 35 is the number of different senses actually appearing in the PDTB 3.0 incl. *+Belief* and *+SpeechAct* aspects.

[5] There is e.g. a *gradation* relation in the Prague taxonomy, prototypically expressed by multi-part *not only... but also* connective).

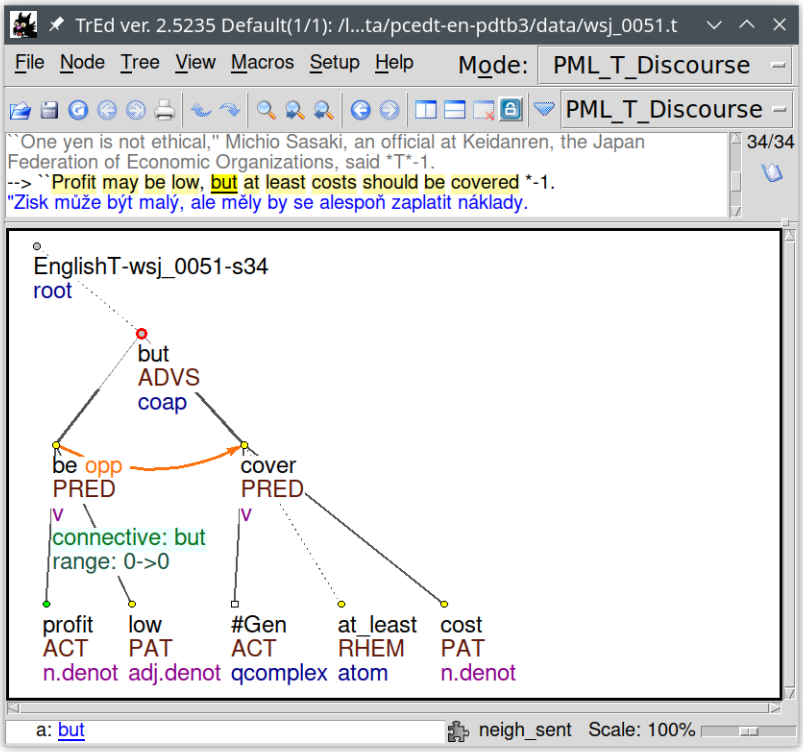


Figure 2. Annotation of the sentence from Example 1 in the Prague discourse annotation tool TrEd

tag, (iii) a surface-syntactic layer (analytical, a-layer): a dependency tree capturing surface syntactic relations such as subject, object, adverbial, (iv) a deep-syntactic layer (tectogrammatical, t-layer): a dependency tree capturing deep syntactic relations (semantically interpreted using labels called functors), ellipses, valency and coreference.

Two major versions of the annotation of discourse relations in the PDT data were published as the Prague Discourse Treebank 1.0 and the Prague Discourse Treebank 2.0. The first version (PDiT 1.0) captured discourse relations marked by explicit connectives (covering conjunctions, adverbs, particles, some types of punctuation marks, some uses of relative pronouns and some types of idiomatic multi-word phrases) and arguments (text units) they connect (Poláková et al., 2013; Mírovský et al., 2014; Zikánová et al., 2015). The data were later updated in PDiT 2.0 (Rysová et al., 2016) with annotation reflecting the division of connectives into *primary* connectives (grammati-

calized single-word units or non-compositional multi-word units) and *secondary* connectives[6] (not yet fully grammaticalized, compositional structures such as *this is the reason why*, *under these conditions*, etc.; Rysová and Rysová, 2014, 2018). In total, there are 21 thousand annotated occurrences of discourse relations expressed by explicit connectives, out of which 20 thousand are expressed by primary connectives.

In contrast with the Penn style, the Prague Discourse Treebank annotation does not include implicit relations, second senses of relations (i.e., always a single sense is attached to a relation), and also attribution is not annotated.[7]

Figure 2 shows the annotation of the discourse relation in the sentence from Example 1 in the Prague style of discourse annotation,[8] displayed in tree editor TrEd[9] (Pajas and Štěpánek, 2008). The discourse relation is expressed by an arrow connecting roots of subtrees corresponding to the arguments of the relation. Its direction indicates the argument semantics (i.e., it corresponds to the third level of senses in the Penn style).

The upcoming Prague Discourse Treebank 3.0 brings a substantial revision of discourse types assignment from the previous release, based in large part on the prior work on the Lexicon of Czech Discourse Connectives (CzeDLex; Mírovský et al., 2021) and, as elaborated and discussed in the rest of the present paper, it offers the annotation of discourse relations also in the Penn style (incl. the Penn sense taxonomy).

3. Transformation of Senses

The transformation process from the Prague style to the Penn style of discourse annotation consists of two separate parts: (i) transformation of the data format, which – although complex – is more a technical than a theoretical problem and we mention it only briefly in Section 3.7, and (ii) transformation of Prague discourse types to Penn senses. The latter brings up a number of theoretical questions that are discussed in the subsequent text.

Table 1 shows a transformation table from Prague discourse types (on the left) to the second level of Penn senses (on the right), based on a detailed study of the annotation manuals and the data of the two corpora. For asymmetric relations, the third level of senses (the argument semantics) is assessed from the direction of the discourse arrow in the Prague annotation.

At a first glance we can make several observations: (i) most discourse types transform to a single sense, (ii) some discourse types transform to two senses, (iii) some

[6] roughly corresponding to *alternative lexicalizations* in the Penn style

[7] More complete discourse annotation, incl. annotation of implicit relations, has been done on a relatively small part of the PDiT data only and published separately as Enriched Discourse Annotation of PDiT Subset 1.0 (PDiT-EDA 1.0; Zikánová et al., 2018).

[8] The underlying tectogrammatical tree comes from the Prague Czech–English Dependency Treebank (PCEDT; Hajič et al., 2012).

[9] `https://ufal.mff.cuni.cz/tred/`

senses correspond to more than one discourse type, and (iv) the division to the four major classes[10] sometimes changes after the transformation.

Observations (iii) and (iv) are not substantial for the present task of Prague to Penn transformation. The Penn senses that correspond to more than one Prague discourse type (e.g., *Comparison.Concession*) merely account in this transformation direction for an (unavoidable) information loss and would only represent an issue for the opposite direction of transformation (Penn to Prague).[11]

Changes in the division to the four main sense classes are a matter of different underlying theoretical categorizations. They take place in such cases of Prague discourse types that were newly introduced for the Prague annotation and did not exist in the PDTB 2.0. The *restrictive opposition* discourse type, for instance, is a wider relation than *Expansion.Exception*, it also encompasses a more relaxed restriction of the content of the other argument. This includes a contrastive (or polarity-change) feature and also contrastive connectives are often used. The affiliation of *correction* and *gradation*, the other two Prague-only labels, to the Comparison class is based on the same principle, cmp. for example the contrastive feature in the complex *not only but also* connective.

On the other hand, observations (i) and (ii) are of the utmost importance. Discourse types that transform to a single sense can be processed without further consideration and represent a fully automatic part of the transformation. Discourse types that transform to two senses need further attention.

Our effort was aimed at discovering to which extent these ambiguous discourse types can be processed automatically with a satisfying success rate and which part of the data needs to be processed manually. The rest of this section is dedicated to a thorough analysis of transformation needs of the individual ambiguous discourse types.

3.1. *Comparison.Similarity* from *conjunction*

One of the relations that is present in the PDTB taxonomy but not in that of PDiT is a relation of *Comparison.Similarity*. *Comparison.Similarity* in the PDTB (Webber et al., 2019) is characterized as follows: "This tag is used when one or more similarities between Arg1 and Arg2 are highlighted with respect to what each argument predicates as a whole or to some entities it mentions."

This sense in PDiT was captured under the relation of *conjunction*. In the preparation of the transformation process, we examined all PDTB occurrences of *Comparison.Similarity* relation and took under scrutiny all connectives used for this sense.

[10] *TEMPORAL*, *CONTINGENCY*, *COMPARISON*, *EXPANSION*

[11] Regarding the sense *Expansion.Level-of-detail* corresponding to two Prague discourse types (*specification* and *generalization*), this ambiguity in the opposite transformation process would easily be solved by taking into account the third level of the PDTB sense hierarchy (argument ordering) and the direction of the relation in the Prague style.

PDiT discourse type	PDTB 3.0 sense(s)
TEMPORAL	
precedence-succession	Temporal.Asynchronous
synchrony	Temporal.Synchronous
CONTINGENCY	
reason-result	Contingency.Cause, Contingency.Negative-cause
pragmatic reason-result	Contingency.Cause+Belief, Contingency.Cause+SpeechAct
condition	Contingency.Condition, Contingency.Negative-condition
pragmatic condition	Contingency.Condition+SpeechAct, Contingency.Negative-condition+SpeechAct
purpose	Contingency.Purpose
explication	Contingency.Cause+Belief
COMPARISON	
confrontation	Comparison.Contrast
opposition	Comparison.Concession
pragmatic contrast	Comparison.Concession+Belief, Comparison.Concession+SpeechAct
restrictive opposition	Expansion.Exception, Comparison.Contrast
concession	Comparison.Concession
correction	Expansion.Substitution
gradation	Expansion.Conjunction
EXPANSION	
conjunction	Expansion.Conjunction, Comparison.Similarity
instantiation	Expansion.Instantiation
specification	Expansion.Level-of-detail
generalization	Expansion.Level-of-detail
equivalence	Expansion.Equivalence
conjunctive alternative	Expansion.Disjunction
disjunctive alternative	Expansion.Disjunction

Table 1. Basic transformation table from PDiT discourse types to the PDTB 3.0 second-level senses

Then we looked for their counterparts in Czech that occurred within the PDiT relation of *conjunction*. In this way, we found *Comparison.Similarity* connectives in Czech, namely single-word connectives *obdobně* [*similarly*] and *podobně* [*similarly*], and their complex variants such as *podobně i* [*similarly also*] or *podobně jako* [*similarly as*], as well as complex connectives containing the word *stejně* [*equally, still*] such as *stejně tak* or *stejně jako* [both meaning *likewise*], see Example 2.

(2) *Také v tomto případě jde o autonomní aktivitu finanční instituce, nota bene na vládě nezávislé, zmocněné k tomu zákonem.* Podobně **vláda využívá mzdové regulace, vyžaduje-li to nárůst inflace.** (PDiT, ln94200_126)

[*Also in this case, it is an autonomous activity of a financial institution, nota bene independent of the government, authorized to do so by law.* Similarly, **the government uses wage regulation if inflation increases.**]

The one-word connective *stejně* [*equally, still, anyway*] did not appear expressing the relation of *Comparison.Similarity*, and therefore it was not covered within this sense, see Example 3.

(3) *Demokracii si můžeme dovolit, protože máme nejlepší a historicky spravedlivý program a národ nás miluje.* **O moc** stejně **nepřijdeme, protože volby vyhrajeme.** (PDiT, ln95048_117)

[*We can afford democracy because we have the best and historically just program and the nation loves us.* **We won't lose power** anyway, **because we will win the elections.**]

3.2. *Contingency.Negative-condition* from *condition*

Another relation that required a deep analysis was the relation of *Contingency.Negative-condition*. In the PDTB manual (Webber et al., 2019), this relation is defined as follows: "This tag is used when one argument (the antecedent) describes a situation presented as unrealized, which if it doesn't occur, would lead to the situation described by the other argument (the consequent). There are distinct senses for interpreting the arguments in terms of semantics or speech acts, with the default being semantics. The label *Contingency.Negative-condition.Arg1-as-negCond* is used when Arg1 describes the antecedent and Arg2, the consequent."

In the analysis of *Contingency.Negative-condition* annotated for English, we focused especially on specific connectives used for this relation and we searched for their counterparts in Czech. We found the following connectives in Czech that were originally annotated as a pure *condition* in PDiT: *jinak* [a counterpart of English *otherwise* and *lest*], *nebo* or *buď_nebo* [counterparts of English *or* and *either_or*] and *aniž* [a counterpart of English constructions containing *without*].

The most challenging situation appeared to be with the connective *unless* (the most frequent connective for *Contingency.Negative-condition* in the PDTB). Czech language

does not have a direct counterpart for this English connective. Thus we faced a complicated issue of how to find Czech contexts in PDiT that correspond meaningfully to English contexts with the connective *unless*.

The connective *unless* contains negation in its sense, but it does not simply mean "if not". However, the presence of negation in the Czech sentence was a basic condition for the search of Czech counterparts of English sentences with *unless*.

The reliable cases that could be marked as *Contingency.Negative-condition* automatically were those in which a connective expressing discourse type *condition* (*pokud*, *když*, *-li* [all meaning *if*]) and a connective such as *tedy* [*that is*], *ovšem* or *však* [both meaning *however*] occurred together in the sentence containing a negation, see Example 4.

(4) *Za rok tu jsem znova*, tedy pokud **mě nepřejede auto.** (PDiT, ln94207_54)
[*I'll be here again in a year* unless **I get run over by a car.**]

However, the second connective (like *tedy* [*that is*] in the example) occurs explicitly in the sentence rather rarely. Therefore, we were looking for other tendencies that characterize the relation of *Contingency.Negative-condition* in Czech.

It turned out that these are the order of the discourse arguments in combination with a particular connective. A big portion of cases that were evaluated as *Contingency.Negative-condition* contained a connective *pokud* or *-li* [both meaning *if*] in the second argument, see Examples 5 and 6.

(5) *Celý rok jsme přečkali bez změny ceny,* **nepočítáme-li zvýšení v souvislosti se zařazením barevného televizního magazínu Duha jako přílohy LN**. (PDiT, ln94210_111)
[*We went the whole year without a price change* unless **we count the increase in connection with the inclusion of the color TV magazine Duha as a supplement to LN**.]

(6) *Mělo by to stačit*, pokud **se nevynoří něco nenadálého**. (PDiT, ln94205_130)
[*That should be enough* unless **something unexpected comes up**.]

3.3. *Contingency.Negative-cause* from *reason–result*

Special attention also had to be paid to the relation of *Contingency.Negative-cause.negResult*. According to the PDTB 3.0 manual, this relation "is used when Arg1 gives the reason, explanation or justification that prevents the effect mentioned in Arg2." It also mentions that the relation "was specifically introduced for the lexico-syntactic construction 'too X to Y'."

This construction corresponds to Czech complex connectives *na to, aby* or *k tomu, aby* that occur together with an adjunct expressing manner by specifying extent or intensity of the event or a circumstance, such as *příliš* [*too* (*much*)], see Example 7.

(7) *Jsem* příliš *mladý* na to, abych **žil se založenýma rukama**. (PDiT, mf920925_120)
[*I'm* too *young* to **live with folded hands**.]

These cases were annotated as a relation of *reason–result* in PDiT. However, all of them have been provided with a comment by an annotator that these constructions are rather specific and require further attention. In this regard, the annotation of these cases as *Contingency.Negative-cause.negResult* provides an effective solution also for Czech.

All these cases have a dependent clause labelled on the underlying tectogrammatical layer by the AIM functor[12] and these cases were a part of discourse annotation. To be sure that all such constructions were treated the same way, we searched for them also in compound sentences with a dependent clause labelled with RESL functor,[13] which was originally omitted from the discourse annotation, because a vast majority of RESL clauses do not have a discourse interpretation. In this way, three additional cases were found to be interpreted as *Contingency.Negative-cause.negResult* (and *reason–result* in the Prague taxonomy).

3.4. *Comparison.Contrast* from *restrictive opposition*

Another issue to be solved concerned the relation of *restrictive opposition*. *Restrictive opposition* in the Prague style is a relation in which the validity of the first argument is limited by the content of the second argument or the second argument expresses an exception to the first one (see the PDiT annotation manual, Poláková et al., 2012). So, the scope of the relation is wider than the one of the *Expansion.Exception* PDTB sense.

We primarily converted Prague relations of *restrictive opposition* to the PDTB 3.0 *Expansion.Exception*[14] but sometimes also to *Comparison.Contrast*.[15] We assumed the relation of *Comparison.Contrast* in cases where *restrictive opposition* was not accompanied by the use of a functor RESTR[16] on the underlying tectogrammatical layer.

Firstly, we manually evaluated cases of intra-sentential relations of *restrictive opposition* in a complex sentence in which the subordinate clause did not contain the

[12] This label is used for non-obligatory modifications that express purpose, the intended result or the aim (Mikulová et al., 2005).

[13] This label is used for a non-obligatory modification that "expresses manner by specifying the result of the event" (Mikulová et al., 2005).

[14] "This tag is used when one argument evokes a set of circumstances in which the described situation holds, and the other argument indicates one or more instances where it doesn't," see the PDTB 3.0 manual (Webber et al., 2019).

[15] "Contrast is used when at least two differences between Arg1 and Arg2 are highlighted," see the PDTB 3.0 manual (Webber et al., 2019).

[16] Label RESTR (restriction) is used for a non-obligatory modification that "expresses manner by specifying an exception/restriction" (Mikulová et al., 2005).

functor RESTR. We found out that the most cases of *Comparison.Contrast* appeared in sentences with connectives *však* [*however*] and (*i*) *když* [*although*].

In the next step, we thus limited our analysis to these connectives and extended the search also to inter-sentential relations. We found altogether 114 occurrences of such a type of sentence and manually marked 86 of them as a relation of *Comparison.Contrast*, see Example 8.

(8) *Lidé na všech stupních řízení jsou schopní, mají snahu se dále učit*. **Chybí jim** však **zkušenosti z dlouhodobého působení**. (PDiT, cmpr9410_010)

[*People at all levels of management are efficient and eager to learn*. However, **they lack long-term experience**.]

The rest of these sentences were annotated as *Expansion.Exception*, see Example 9.

(9) *Jeho návrh hovoří o šecích, které by následně získaly domácnosti od státu na placení všech faktur za energie, které domácnost využije*. **Vyloučeny by** však **byly motorové kapalné pohonné hmoty**. (PDiT, cmpr9410_049)

[*His proposal talks about checks that households would subsequently receive from the state to pay all invoices for energy that the household uses*. However, **liquid motor fuels would be excluded**.]

3.5. Pragmatic Relations

Three pragmatic relations were established in the Prague taxonomy of discourse types – namely *pragmatic reason–result*, *pragmatic condition* and *pragmatic contrast*. Although these relations were originally inspired by the PDTB 2.0 pragmatic relations, they were in the Prague style defined broader: these labels were used for cases where the semantics and the form do not correspond to each other. In a vast majority of cases, such a relation holds between one argument and a content that is inferred from the other argument. Analysis of all pragmatic relations in PDiT (Poláková and Synková, 2021) showed that this discrepancy/inference can be of various kinds, two of them corresponding to PDTB 3.0 relations with *+Belief* and *+SpeechAct* aspects (namely *Contingency.Cause+Belief*, *Contingency.Cause+SpeechAct*, *Contingency.Condition+SpeechAct* and *Comparison.Concession+SpeechAct*). *Contingency.Cause+Belief* "is used when evidence is provided to cause the hearer to believe a claim. The belief is implicit." (PDTB 3.0 manual; Webber et al., 2019), tags with *+SpeechAct* aspect were used when a relation holds between an argument and an implicit speech act represented by the other argument (PDTB 3.0 manual) – see Example 10.

(10) Jestliže **chcete slyšet můj postoj k rozhodnutí poroty**, *je to neslýchaný projev neúcty k práci druhého*. (PDiT, lnd94103_102)

[If **you want to hear my take on the jury's decision**, *it's an unheard of disrespect for someone else's work.*]

In contrast to the Penn definition, *pragmatic reason–result* relations in the Prague style corresponding to the Penn relation of *Contingency.Cause+Belief* have also the subjectivity aspect – a claim or provided evidence was a highly subjective one, as showed by Example 11.

(11) *Nemají se za co omlouvat, ale zároveň se nesmějí starat jen o sebe a svá konta*. Proto **by měli deset procent z vyhraných peněz věnovat na charitu**. (PDiT, ln94208_106)

[*They have nothing to apologize for, but at the same time they must not only care about themselves and their accounts*. Therefore, **they should donate ten percent of the money won to charity**.]

Besides these relations (corresponding to Penn *+Belief* and *+SpeechAct* relations), there were also cases where pragmatic relations in PDiT were annotated because of a complicated inference resulting from a cultural context, and cases with broken coherence caused by a formulation clumsiness. These relations were transformed to Penn senses without the *+Belief* and *+SpeechAct* aspects.

Discourse types of all pragmatic relations in PDiT were transformed to the corresponding Penn senses manually because there is no formal clue for distinguishing cases with *+Belief* and *+SpeechAct* aspects, and cases without them.

Altogether, 35 of 100 pragmatic relations in PDiT were transformed to relations with *+Belief* or *+SpeechAct* aspects, leaving the rest of them labelled as *Contingency.Cause*, *Contingency.Condition* or *Comparison.Concession*.

The above analysis has shown that the relation of *pragmatic condition* in PDiT was annotated quite rarely, implying a possible high number of false negatives. So a probe was performed in the whole data to see if some *pragmatic conditions* were by mistake annotated as *conditions*. As some *pragmatic conditions* were indeed found in the analyzed sample of relations of *condition*, all *condition* relations were then checked manually and 92 *pragmatic conditions* (corresponding to *Contingency.Condition+SpeechAct*) were newly annotated. One of them is given in Example 12.

(12) Kdybych **měl jmenovat konkrétní autory**, *byla by jich spousta*. (PDiT, ln95048_050)

[If **I should name specific authors**, *there would be lots of them*.]

3.6. *Specification* with the List Relation

The Prague annotation style recognizes a special type of relation called *list*. The list relation holds between enumerated items (i.e. *first*, *second*; *1*), *2*) etc.) and these items as a whole are connected with its hypertheme (i.e., sentences such as *there are several problematic issues*) by a *specification* relation that can (contrary to *specification* relation not related to a list) be without a connective or can hold between nominal arguments.

As the list relation does not have a counterpart in the Penn style of annotation,[17] it is omitted from the transformation. However, the introductory *specification* relation has its counterpart in *Expansion.Level-of-detail.Arg2-as-detail* sense, so all *specification* relations connected to a list had to be checked manually to decide which of them can be interpreted also as explicit *Expansion.Level-of-detail.Arg2-as-detail*. From 82 *specification* relations connected with a list relation, 16 cases could be transformed to the corresponding Penn style relation.

3.7. Technical notes

From all technical parts of the transformation process, the extraction of arguments of the relations from their deep-syntactic tree representations to plain text proved to be the most challenging one. The numerous issues can be split in two categories: (i) annotation inconsistencies in various parts of the data (on the deep-syntactic layer, on the surface-syntactic layer, in the discourse annotation), and (ii) a complex nature of the deep-syntactic layer of annotation (reconstructed nodes/parts of the trees that take part in discourse relations, necessity to combine information from several annotation layers). Although we took great care in tuning the plain text generation of the arguments, we could not check and fix errors in all 21 thousand of discourse relations.

To demonstrate the kind of phenomena involved in discourse relations with elided (and reconstructed) nodes, consider Examples 13 and 14.

(13) *... nechtěli* [*povolit*] nebo **nemohli odklad platby povolit** (PDiT, cmpr9410_002)
[*... would not* [*allow*] or **could not allow payment deferral**]

(14) *Celní unie bude existovat na papíře ještě dalších dvanáct měsíců* (a **třeba [bude existovat] i déle**) ... (PDiT, cmpr9410_001)
[*The customs union will exist on paper for another twelve months* (and **maybe [will exist] even longer**) ...]

In both cases, a discourse relation holds technically between two tectogrammatical nodes representing the same content verb, one of them being elided in the surface form of the sentence: *povolit* [*to allow*] in the first example and *existovat* [*to exist*] in the second example. In the first case, the actual discourse relation holds rather between the auxiliary verbs *nechtěli* [*would not*] and *nemohli* [*could not*], and although auxiliary nodes are not directly present at the tectogrammatical layer, they need to be represented in the plain text versions of the arguments. On the contrary, in the second

[17] From the list of implicit connectives – i.e. connectives filled in by annotators when annotating implicit relations – it seems that the Prague type list would be labeled as *Expansion.Conjunction*, because expressions *first*, *second*, *third* are listed there as connectives of implicit *Expansion.Conjunction* relations (PDTB 3.0 manual, Webber et al., 2019). However, expressions *first*, *second*, *third* are not listed in the list of explicit connectives, so this interpretation is just a guess.

case, the auxiliary node *bude* [*will*] needs to be present only in the first argument and omitted from the second one.[18]

Further, in the Prague style of discourse annotation, supplementary text parts were not annotated separately from the argument delimitation. Although the minimality principle was followed, in cases where the surrounding sentences played a distinct role in the discourse relation, they were marked as a part of the argument. In such cases, the additional sentences are transformed to the Penn style as supplementary texts.

Definitions of all data fields in the column format used for the transformed PDiT data are given in Table 3 in Appendix. Most of them come from the PDTB 3.0 data format; we have added a few fields to keep the original Prague discourse type and to provide plain text versions of information only captured in the form of spans in other fields.

4. Results and Discussion

Table 2 represents an overview of the result of the Prague discourse types to Penn senses transformation in the Prague Discourse Treebank data. The table shows a comparison of distributions of senses in (transformed) PDiT 3.0 and the PDTB 3.0 (in the latter taking into account explicit discourse relations only[19]). The two corpora are close to each other in size (both approx. 50 thousand sentences), genres (journalistic texts), in total numbers of explicit discourse relations (21 thousand vs. 25 thousand) and, as can be observed in the table, also in distributions of explicit discourse relations senses.

Although the sense frequencies in the two corpora are close in most of the cases, for several senses there are noticeable differences – they are highlighted in the table with grey background. Some of them may have roots in differences in the theoretical backgrounds of the two annotation styles, some others may simply reflect language or corpora differences. This constitutes a research question which inspired the following analysis. Let us elaborate below on the individual cases of noticeable differences in sense frequencies; for each sense, we state in parentheses the numbers of occurrences in the PDiT 3.0 transformed data and in the PDTB 3.0 data (but considering the slightly different total numbers of explicit relations in the two corpora, please take into account also the relative frequencies in the table).

[18] ...although it is referenced (via a link to the surface-syntactic layer) from both nodes representing the content verb *existovat* [*to exist*]. This can happen even in discourse relations between two non-elided content verbs, e.g. *Trámy byly urychleně rozebrány ...* a **[byly] odvezeny do dílen ...** (PDiT, ln94210_95) [*The beams were quickly disassembled* and **[they were] taken to the workshops ...**].

[19] i.e., in the PDTB terminology, relations marked as Explicit, AltLex and AltLexC

sense	PDiT	%	%	PDTB
Comparison.Concession.Arg1-as-denier	568	2.6%	2.9%	742
Comparison.Concession.Arg2-as-denier	3 551	16.4%	15.7%	4 057
Comparison.Concession+SA.Arg2-as-denier+SA	4	0.0%	0.1%	17
Comparison.Contrast	780	3.6%	4.5%	1 155
Comparison.Similarity	47	0.2%	0.7%	169
Contingency.Cause.Reason	1 750	8.1%	6.6%	1 712
Contingency.Cause.Result	1 299	6.0%	4.5%	1 160
Contingency.Cause+Belief.Reason+Belief	123	0.6%	0.1%	34
Contingency.Cause+Belief.Result+Belief	7	0.0%	0.0%	7
Contingency.Cause+SA.Reason+SA	2	0.0%	0.0%	1
Contingency.Cause+SA.Result+SA	4	0.0%	0.0%	1
Contingency.Condition.Arg1-as-cond	48	0.2%	0.1%	27
Contingency.Condition.Arg2-as-cond	1 237	5.7%	5.6%	1 445
Contingency.Condition+SA	102	0.5%	0.3%	73
Contingency.Negative-cause.NegResult	8	0.0%	0.0%	4
Contingency.Negative-condition.Arg1-as-negCond	2	0.0%	0.1%	16
Contingency.Negative-condition.Arg2-as-negCond	48	0.2%	0.4%	110
Contingency.Purpose.Arg1-as-goal	6	0.0%	0.5%	117
Contingency.Purpose.Arg2-as-goal	415	1.9%	1.2%	299
Expansion.Conjunction	8 161	37.8%	34.4%	8 907
Expansion.Disjunction	367	1.7%	1.2%	304
Expansion.Equivalence	127	0.6%	0.1%	37
Expansion.Exception.Arg1-as-excpt	6	0.0%	0.1%	15
Expansion.Exception.Arg2-as-excpt	195	0.9%	0.1%	24
Expansion.InstantiationArg1-as-instance	2	0.0%	0.0%	3
Expansion.InstantiationArg2-as-instance	206	1.0%	1.4%	375
Expansion.Level-of-detail.Arg1-as-detail	136	0.6%	0.2%	51
Expansion.Level-of-detail.Arg2-as-detail	646	3.0%	1.0%	262
Expansion.Manner.Arg1-as-manner	-	-	0.0%	3
Expansion.Manner.Arg2-as-manner	-	-	1.1%	280
Expansion.Substitution.Arg1-as-subst	61	0.3%	0.4%	111
Expansion.Substitution.Arg2-as-subst	391	1.8%	0.5%	137
Temporal.Asynchronous.Precedence	686	3.2%	4.1%	1 071
Temporal.Asynchronous.Succession	341	1.6%	4.5%	1 171
Temporal.Synchronous	262	1.2%	7.7%	1 981
total	21 588	100%	100%	25 878

Table 2. Comparison of distributions of senses in PDiT 3.0 and the PDTB 3.0. Please note that in the names of the senses, 'SpeechAct' was shortened to 'SA' to fit the page. Substantially different frequencies are highlighted with grey background.

Comparison.Similarity (47 in PDiT vs. 169 in the PDTB)

This difference results from different theoretical decisions. In the Prague style, all dependent clauses expressing manner were left out of the annotation, because they were considered not to be a separate abstract object and therefore did not form a discourse argument. Manner can be expressed also by comparison – and similarity is one type of comparison. Thus all cases of *Comparison.Similarity* in transformed PDiT come from discourse type of *conjunction* and do not appear in constructions with a dependent clause expressing manner by means of comparison.

Contingency.Cause+Belief.Reason+Belief (123 vs. 34)

Difference in the frequencies of this relation lies in our opinion in the fact that Czech has a special connective signalling this relation, connective *totiž* [*you see, actually*], which, besides other functions, can signal an argument for a claim. All examples of *Contingency.Cause+Belief* in the PDTB 3.0 manual (Webber et al., 2019) use details (not a reason) as evidence of justification for the presented claim and a majority of them are implicit (without a connective); in Czech, connective *totiž* is used in such contexts. Relations with this connective form 50 percent of all instances of *Contingency.Cause+Belief.Reason+Belief* in PDiT.

Contingency.Purpose.Arg1-as-goal (6 vs. 117)

Except for two cases, all instances of this relation in the PDTB 3.0 have connective *by* which can be in Czech expressed either by a dependent clause with connective *tím, že* [lit. *by that that*] or by a noun in the instrumental case (i.e. without any conjunction or preposition, without any connective) – none of these options is considered to be discourse relevant in the Prague style. Besides, these relations in the PDTB 3.0 hold mostly between arguments without finite verbs – as shown by Example 15. So this difference reflects both theoretical and language differences.

(15) *to correct this problem* by **providing a reliable flow of lendable funds** (PDTB, wsj_1131)

Expansion.Equivalence (127 vs. 37)

We could not find a satisfactory explanation for the different frequencies of this relation. It may be given by the polysemous nature of connective *tedy*, which corresponds to English *so*, *therefore*, but also to connective *in other words* and in some contexts more interpretations are possible. *Expansion.Equivalence* relations with connective *tedy* form 40 percent of all instances of this relation in PDiT.

Expansion.Exception.Arg2-as-excpt (195 vs. 24)

As described in detail in section 3.4, the PDiT relation of *restrictive opposition* corresponds partially to *Expansion.Exception* and at the same time includes also cases which would be interpreted as *Comparison.Contrast* in the PDTB 3.0 taxonomy. Manual analysis of the *restrictive opposition* relation in PDiT covered only the most frequent constructions and connectives, not all instances of the relation.

Expansion.Level-of-detail.Arg2-as-detail (646 vs. 262)

This difference stems from a theoretical decision to consider a colon and a dash to be discourse connectives in the Prague style – relations *Expansion.Level-of-detail.Arg2-as-detail* with these connectives form 60 percent of all instances of this relation in PDiT.

Expansion.Manner (0 vs. 283)

As already mentioned above, in the Prague style, clauses expressing manner were not considered to be separate abstract objects, so they were treated as a syntactic, not a discourse phenomenon.

Expansion.Substitution.Arg2-as-subst (391 vs. 137)

The higher frequency of this relation in PDiT is in our opinion given by the nature of the underlying PDT data – namely by the fact that elided verbs are reconstructed in the dependency trees of the deep-syntactic (tectogrammatical) layer of the corpus, thus allowing to annotate discourse relations with two verbal arguments in constructions such as *it is not A but B* (in Czech typically with an elided verb in B). For example, in the second part of the context in Example 16, there is a node for elided verb *poskytnout* [*to provide*]. Reconstructed nodes for elided verbs take part in annotation of 40 percent of all relations *Expansion.Substitution.Arg2-as-subst* in PDiT.

(16) *Tyto prostředky neposkytne místním spotřebitelům*, ale **[poskytne je] japonským zemědělcům.** (PDiT, ln94208_147)
[*It will not provide these funds to local consumers*, but **[it will provide them] to Japanese farmers.**]

Temporal.Synchronous (262 vs. 1981)

Upon close examination, we attribute this difference to a large extent to theoretical differences in the two annotation styles. Frequencies of translation counterparts of the most common connectives with this sense differ substantially. For example, whereas the most frequent Czech connective for this sense *když* has 783 occurrences in PDiT and only 100 of them are annotated as *Temporal.Synchronous*, its English counterpart

when has 1 076 occurrences in the PDTB and half of them are assigned the *Temporal.Synchronous* sense (Webber et al., 2019). Besides, approx. 650 of the PDTB 3.0 *Temporal.Synchronous* relations have been labelled also by a second sense (*Comparison.Contrast*, *Contingency.Cause.Reason* etc.). As second senses are not annotated in the Prague style, sometimes other discourse types than temporal took precedence in the PDiT annotation if they were present in the given context. In contexts such as in Example 17, the Prague style would annotate just the *reason–result* relation, whereas the Penn style annotates *Temporal.Synchronous* as the first sense and *Contingency.Cause.Reason* as the second sense.

(17) *The company acquired the debt* when **it paid $155 million to purchase Wilson last year** (PDTB, wsj_0510)

5. Conclusion

The Prague Discourse Treebank data transformed to the Penn style of discourse annotation was published in December of 2022 in LINDAT/CLARIAH-CZ repository under the Creative Commons licence[20] as the Prague Discourse Treebank 3.0 (PDiT 3.0; Synková et al., 2022). The data was published in two formats: (i) the original Prague format of discourse annotation on top of tectogrammatical trees,[21] and (ii) the Penn column format of discourse annotation accompanied by the original plain texts. The discourse research community thus gets to its disposal another large-scale corpus manually annotated with discourse relations in the PDTB 3.0 style.

Understanding of the differences between the Prague and Penn semantic types taxonomies and of limits of the automatic transformation of the Prague discourse types to the Penn senses, based on a detailed study of both respective corpora, their annotation manuals and on a comparison of distributions of discourse relation senses in the two corpora, belong to the main theoretical results of the presented research. Frequencies of senses in the transformed PDiT data and in the PDTB 3.0 data are interestingly very similar. We have discussed the cases of senses where these frequencies considerably differed.

Differences in the taxonomies may in some cases reflect differences in the languages. For example, English has a particular connective – *unless* – for the relation of *Contingency.Negative-condition*, while Czech does not have its direct counterpart. During the conversion of the PDiT discourse annotation to the Penn style, we encountered a need to take a deeper look at how sentences corresponding to the English usage of

[20] `http://hdl.handle.net/11234/1-4875`

[21] For licensing reasons, the PDiT 3.0 distribution does not actually contain the tectogrammatical trees (and the lower layers of annotation); instead, the underlying data needs to be downloaded separately from the LINDAT/CLARIAH-CZ repository (the PDT part of the PDT-C 1.0, `http://hdl.handle.net/11234/1-3185`) and the discourse annotation can be added to the data by a script provided by the PDiT 3.0 distribution.

unless are constructed in Czech. In this way, we found certain tendencies combining the use of a particular connective, sentence negation and the position of discourse arguments.

The theoretical results are reflected also in technical procedures developed during the presented research for transforming the Prague style of discourse annotation to the Penn style. These procedures can be used in future for any data annotated in the Prague style of discourse annotation. They consist of two separate parts: (i) transformation of discourse arguments and connectives from their representation in tectogrammatical trees to plain text, and (ii) transformation of Prague discourse types to Penn senses.

Thousands of discourse relations in the PDiT data were examined during the research, resulting in many rules embedded in the transformation procedures. These rules were used to transform discourse types of 54 percent of all PDiT discourse relations (12 thousand out of over 21 thousand). 42 percent (over 9 thousand) of the PDiT relations carry a discouse type that transforms to a single Penn sense; their discourse types were also transformed automatically. In the end, discourse types of only 1.8 percent of all discourse relations in the PDiT data (388 relations) had to be disambiguated manually in order to be transformed to the correct sense.

The project covering this research will continue for two more years, having as its ultimate goal to have the whole Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)[22] annotated with discourse relations and published in both the Prague and Penn styles of discourse relations annotation.

Acknowledgement

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project 22-03269S). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2018101).

Bibliography

Bojar, Ondřej, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. Evald Reference-less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 541–545, 2018. doi: 10.18653/v1/W18-6432.

Danlos, Laurence, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. Vers le FDTB: French Discourse Tree Bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 471–478, 2012.

[22] The Prague Dependency Treebank - Consolidated 1.0 consists of four subcorpora: (i) the PDT, (ii) the Czech part of the Prague Czech-English Dependency Treebank, (iii) the Prague Dependency Treebank of Spoken Czech, and (iv) Faust. Altogether, PDT-C 1.0 includes approx. 175 thousand sentences.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of NAACL: Human Language Technologies*, pages 4171–4186, 2019.

Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0), 2020.

Hajič, Jan, 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. *Prague Czech-English Dependency Treebank 2.0*. Data/Software, Linguistic Data Consortium, 2012. University of Pennsylvania, Philadelphia. LDC2012T08.

Kiddon, Chloé, Luke Zettlemoyer, and Yejin Choi. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, 2016. doi: 10.18653/v1/D16-1032.

Lee, Alan, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Annotating Discourse Relations with the PDTB Annotator. In *Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: System Demonstrations*, pages 121–125, 2016.

Mann, William C. and Sandra A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281, 1988. doi: 10.1515/text.1.1988.8.3.243.

Meyer, Thomas and Bonnie Webber. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, 2013.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Technical report, ÚFAL MFF UK, Prague, 2005.

Mírovský, Jiří and Lucie Poláková. Sense Prediction for Explicit Discourse Relations with BERT. In Yang, Xin-She, Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Sixth International Congress on Information and Communication Technology (ICICT)*, volume 216 of *Lecture Notes in Networks and Systems*, pages 835–842, Singapore, 2021. International Congress and Excellence Awards, Springer. ISBN 978-981-16-1781-2.

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. Does Tectogrammatics Help the Annotation of Discourse? In *Proceedings of COLING 2012: Posters*, pages 853–862, 2012. URL `https://www.aclweb.org/anthology/C12-2083.pdf`.

Mírovský, Jiří, Pavlína Jínová, and Lucie Poláková. Discourse Relations in the Prague Dependency Treebank 3.0. In Tounsi, Lamia and Rafal Rak, editors, *The 25th International Conference on Computational Linguistics* (*Coling 2014*), *Proceedings of the Conference System Demonstrations*, pages 34–38, Dublin, Ireland, 2014. Dublin City University (DCU), Dublin City University (DCU).

Mírovský, Jiří, Pavlína Synková, Lucie Poláková, Věra Kloudová, and Magdaléna Rysová. CzeDLex 1.0, 2021.

Oza, Umangi, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. The Hindi Discourse Relation Bank. In *Proceedings of the third Linguistic Annotation Workshop*, pages 158–161, 2009. doi: 10.3115/1698381.1698410.

Pajas, Petr and Jan Štěpánek. Recent Advances in a Feature-rich Framework for Treebank Annotation. In Scott, Donia and Hans Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester, 2008. The Coling 2008 Organizing Committee. doi: 10.3115/1599081.1599166. URL https://www.aclweb.org/anthology/C08-1085.pdf.

Poláková, Lucie and Pavlína Synková. Pragmatické aspekty v popisu textové koherence. *Naše řeč*, 104(4):225–242, 2021. ISSN 0027-8203.

Poláková, Lucie, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, and Eva Hajičová. Manual for Annotation of Discourse Relations in Prague Dependency Treebank. Technical Report 47, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2012.

Poláková, Lucie, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, 2013. Asian Federation of Natural Language Processing.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC'08*), pages 2961–2968, Marrakech, 2008. European Language Resources Association. URL http://lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

Prasad, Rashmi, Bonnie Webber, Alan Lee, and Aravind Joshi. *Penn Discourse Treebank Version 3.0*. Data/Software, Linguistic Data Consortium, 2019. URL https://catalog.ldc.upenn.edu/LDC2019T05. University of Pennsylvania, Philadelphia. LDC2019T05.

Rysová, Kateřina, Magdaléna Rysová, and Jiří Mírovskỳ. Automatic Evaluation of Surface Coherence in L2 Texts in Czech. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing* (*ROCLING 2016*), pages 214–228, 2016.

Rysová, Magdaléna and Kateřina Rysová. The Centre and Periphery of Discourse Connectives. In *Proceedings of Pacific Asia Conference on Language, Information and Computing*, pages 452–459, Bangkok, 2014. Department of Linguistics, Faculty of Arts, Chulalongkorn University. URL https://www.aclweb.org/anthology/Y14-1052.pdf.

Rysová, Magdaléna and Kateřina Rysová. Primary and secondary discourse connectives: Constraints and preferences. *Journal of Pragmatics*, 130:16–32, 2018. ISSN 0378-2166. doi: 10.1016/j.pragma.2018.03.013.

Rysová, Magdaléna, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, and Šárka Zikánová. Prague Discourse Treebank 2.0. Data/Software. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2016. URL `http://hdl.handle.net/11234/1-1905`.

Shi, Wei and Vera Demberg. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5794–5800, 2019. doi: 10.18653/v1/D19-1586.

Synková, Pavlína, Magdaléna Rysová, Jiří Mírovský, Lucie Poláková, Veronika Sheller, Jana Zdeňková, Šárka Zikánová, and Eva Hajičová. Prague Discourse Treebank 3.0, 2022.

Taboada, Maite and William C Mann. Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse studies*, 8(3):423–459, 2006. doi: 10.1177/1461445606061881.

Turney, Peter D and Michael L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003. doi: 10.1145/944012.944013.

Webber, Bonnie, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545–587, 2003. doi: 10.1162/089120103322753347.

Webber, Bonnie, Rashmi Prasad, Alan Lee, and Aravind Joshi. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*, 35:108, 2019.

Xiong, Hao, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling Coherence for Discourse Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345, 2019. doi: 10.1609/aaai.v33i01.33017338.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 1–16, 2015. doi: 10.18653/v1/K15-2001.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proc. of the CoNLL-16 shared task*, pages 1–19, 2016. doi: 10.18653/v1/K16-2001.

Zeyrek, Deniz and Murathan Kurfalı. TDB 1.1: Extensions on Turkish Discourse Bank. *LAW XI 2017*, page 76, 2017. doi: 10.18653/v1/W17-0809.

Zhang, Renxian. Sentence Ordering Driven by Local and Global Coherence for Summary Generation. In *Proceedings of the ACL 2011 Student Session*, pages 6–11, 2011.

Zhou, Yuping and Nianwen Xue. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the ACL: Long Papers-Volume 1*, pages 69–77, 2012.

Zikánová, Šárka, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia, 2015. ISBN 978-80-904571-8-8.

Zikánová, Šárka, Pavlína Synková, and Jiří Mírovský. Enriched Discourse Annotation of PDiT Subset 1.0 (PDiT-EDA 1.0), 2018.

Appendix

Index	Field Name	Description
0	Relation Type	Explicit, AltLex, AltLexC
1	Conn SpanList	SpanList of the Explicit Connective or the AltLex/AltLexC selection
2	Conn Src	Connective's Source
3	Conn Type	Connective's Type
4	Conn Pol	Connective's Polarity
5	Conn Det	Connective's Determinacy
6	Conn Feat SpanList	Connective's Feature SpanList
7	Conn1	Explicit Connective Head
8	SClass1A	Semantic Class of the Connective
9	SClass1B	Second Semantic Class of the First Connective
10	Conn2	Second Implicit Connective
11	SClass2A	First Semantic Class of the Second Connective
12	SClass2B	Second Semantic Class of the Second Connective
13	Sup1 SpanList	SpanList of the First Argument's Supplement
14	Arg1 SpanList	SpanList of the First Argument
15	Arg1 Src	First Argument's Source
16	Arg1 Type	First Argument's Type
17	Arg1 Pol	First Argument's Polarity
18	Arg1 Det	First Argument's Determinacy
19	Arg1 Feat SpanList	SpanList of the First Argument's Feature
20	Arg2 SpanList	SpanList of the Second Argument
21	Arg2 Src	Second Argument's Source
22	Arg2 Type	Second Argument's Type
23	Arg2 Pol	Second Argument's Polarity
24	Arg2 Det	Second Argument's Determinacy
25	Arg2 Feat SpanList	SpanList of the Second Argument's Feature
26	Sup2 SpanList	SpanList of the Second Argument's Supplement
27	Adju Reason	The Adjudication Reason
28	Adju Disagr	The type of the Adjudication disagreement
29	PB Role	The PropBank role of the PropBank verb
30	PB Verb	The PropBank verb of the main clause of this relation
31	Offset	The Conn SpanList of Explicit/AltLex/AltLexC tokens
32	Provenance	Indicates whether the token is a new PDTB3 token
33	Link	The link id of the token
34	Discourse Type	The original discourse type in the Prague taxonomy
35	Conn Text	Text representation of field 31 (Offset)
36	Conn Feat Text	Text representation of field 6 (Conn Feat SpanList)
37	Sup1 Text	Text representation of field 13 (Sup1 SpanList)
38	Arg1 Text	Text representation of field 14 (Arg1 SpanList)
39	Arg1 Feat Text	Text representation of field 19 (Arg1 Feat SpanList)
40	Arg2 Text	Text representation of field 20 (Arg2 SpanList)
41	Arg2 Feat Text	Text representation of field 25 (Arg2 Feat SpanList)
42	Sup2 Text	Text representation of field 26 (Sup2 SpanList)
43	Genre	The genre of the document

Table 3. Field definitions in PDiT 3.0 corresponding to fields defined in the PDTB 3.0 (fields 0–33) and additional fields (34–43) present in the PDiT 3.0 column data format. Fields not used in PDiT 3.0 are highlighted with grey background.

Besides the original PDiT format of the data, the transformed discourse annotation is also provided in the PDTB 3.0 column text format where each discourse relation is represented by a single line consisting of a number of fields separated with '|', with each field carrying a single piece of annotation information. For compatibility reasons, we have kept all field definitions from the PDTB 3.0 (although not all of them are actually used in the transformed PDiT data[23]) and for additional information, we have added new fields. Table 3 gives field definitions of the format used for the PDiT transformed data. The first part of the table, fields 0–33, corresponds to the original PDTB 3.0 fields; it is taken from the PDTB 3.0 annotation manual (Webber et al., 2019) and the definitions are adjusted to better fit our data. The second part, fields 34–43, gives definitions of additional fields introduced in the PDiT 3.0 transformed data.

Address for correspondence:
Jiří Mírovský
mirovsky@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic

[23] Neither in the PDTB 3.0 are all of them used, as the PDTB 3.0 keeps backward format compatibility with its previous version.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 120 APRIL 2023 31–46

Universal Dependencies for Malayalam

Abishek Stephen, Daniel Zeman

ÚFAL, Faculty of Mathematics and Physics, Charles University

Abstract

Treebanks can play a crucial role in developing natural language processing systems and to have a gold-standard treebank data it becomes necessary to adopt a uniform framework for the annotations. Universal Dependencies (UD) aims to develop cross-linguistically consistent annotations for the world's languages. The current paper presents the essential pivots of a UD-based syntactically annotated treebank for Malayalam. Sentences extracted from the IndicCorp corpus were manually annotated for morphological features and dependency relations. Language-specific properties are discussed which shed light on many of the grammatical areas in the Dravidian language syntax which needs to be examined in depth. This paper also discusses some pertaining issues in UD taking into consideration the Dravidian languages and provides insights for further improvements in the existing treebanks.

1. Introduction

A treebank is a collection of syntactically and (or) semantically annotated language data. Most treebanks are developed using a combination of manual and automated processes (Kakkonen, 2006). Many treebanks based on dependency grammar (Jiang and Liu, 2015; de Marneffe and Nivre, 2019) have been developed recently. The Prague Dependency Treebank (Hajič et al., 2020) of Czech is one of the largest dependency treebanks. But with the growing need for multilingual language systems and better cross-linguistic evaluations, a uniform framework is needed. Universal Dependencies (UD) (de Marneffe et al., 2021) is a framework for consistent annotation of natural language data (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with

 Corresponding author: stephen@ufal.mff.cuni.cz
Cite as: Abishek Stephen, Daniel Zeman. Universal Dependencies for Malayalam. The Prague Bulletin of Mathematical Linguistics No. 120, 2023, pp. 31–46. doi: 10.14712/00326585.026.

over 500 contributors who have produced 243 treebanks in 138 languages so far.[1] Currently, UD has treebanks for 10 Indian languages among which there are 3 Dravidian languages including Malayalam, the others being Telugu and Tamil. The Malayalam treebank[2] is a step forward to do better comparative evaluation of syntactic properties of the Dravidian languages and also with other unrelated languages. The following sections of the paper describe how the treebank was developed elaborating on the challenges and the resorts taken.

2. Malayalam

Malayalam is a Dravidian language spoken primarily in the south-Indian state of Kerala. Malayalam is an agglutinating language like the other Dravidian languages. There are 35 million native-speakers of Malayalam in India. Malayalam has borrowed freely from other languages, especially from Sanskrit. That accounts for lemmas of many content words resembling those in Sanskrit. The canonical word order in Malayalam is SOV. Unlike Tamil or Telugu, Malayalam lacks verbal agreement, i.e., the verb does not encode the person, number and gender of the subject (nor those of object or any other argument). We have a three-way distinction of tense in Malayalam, i.e., present, past and future. Additionally, Malayalam has perfective and imperfective aspects along with a number of different moods. Nominalized verb forms are very frequent and so are cleft constructions. Core arguments are marked by the morphological cases nominative (subject) and accusative (object). Core arguments are bare noun phrases without adpositions. Subjects are suppressed when verbs are passivized.

3. Data

The first 20 annotated sentences are Malayalam equivalents of the examples from the Cairo CICLing Corpus.[3] With a preference for texts from different genres in order to get hold of different and unique syntactic constructions, the rest is taken from the Malayalam part of IndicCorp (Kakwani et al., 2020). IndicCorp is a freely available corpus for Indian languages, developed by scraping of web sources comprising of news articles, magazines and books. The corpus contains a single large text file with automatic sentence segmentation, one sentence per line. The publicly released version is randomly shuffled and untokenized.[4] The size of the Malayalam part of IndicCorp exceeds 50 million sentences. The Malayalam treebank currently contains 218 sentences / 2403 words, to be released in UD 2.12.

[1] `https://universaldependencies.org/`

[2] Currently just a small sample of Malayalam grammatical examples.

[3] `https://github.com/UniversalDependencies/cairo`

[4] Available at `https://ai4bharat.iitm.ac.in/corpora`.

4. Methodology

We process the sentences in small batches. After the initial batches, annotation guidelines specific to Malayalam are refined depending on phenomena encountered. After each batch we also retrain a model for tagging and parsing and use it to pre-annotate the next batch, which is then manually corrected in two steps: First, the annotator verifies the annotation of every word including its attachment in the dependency tree, and modifies the annotation where needed. Second, automatic tools are employed to identify errors and inconsistencies, which are then manually corrected. We do not have at our disposal multiple Malayalam-speaking annotators who could annotate the same span and then compare the results. Script-based quality checking should at least partially compensate for this shortcoming.

Manual annotation (including corrections of tokenization and occasionally sentence segmentation) is done in the CoNLL-U Editor (Heinecke, 2019).

4.1. Preprocessing

Unicode NFC normalization is applied to all input sentences. For example, some texts represent the long *ō* (`MALAYALAM VOWEL SIGN OO`, `U+D4B`) as the sequence of *ē* (`MALAYALAM VOWEL SIGN EE`, `U+D47`) and *ā* (`MALAYALAM VOWEL SIGN AA`, `U+D3E`); both representations result in the same glyph. The normalization makes sure to convert them to `U+D4B`, which is the canonical representation. In addition to NFC, we also normalize a few sequences that are used as an alternative representation of so-called chillu letters. These letters are specific syllable-closing variants of certain consonants and they do not have analogy in other Indian scripts. The alternative encoding uses a standard consonant followed by viram (`U+D4D`) and `ZERO WIDTH JOINER` (`U+200D`); we convert any such sequence to the Unicode point dedicated to the resulting chillu consonant.

Furthermore, we generate sentence-level English translation with the help of Google Translate and we use a script[5] to add Latin transliteration of whole sentences as well as of individual word forms. This step is repeated after annotation to also provide transliteration of lemmas.

4.2. Tokenization

In Malayalam, words are delimited by whitespace characters or punctuation. Multiword tokens are relatively common in Malayalam. In the following situations, we understand orthographic tokens as corresponding to multiple syntactic words and split them:

- The copula ആക്കുക (*āk*) 'to be' is written as a suffix of the nominal or adjectival predicate. However, sometimes it is suffixed to another word in the clause, in-

[5] `https://github.com/dan-zeman/translit/blob/main/conllu_translit.pl`

dicating that it is a clitic rather than a derivational morpheme that would derive a verb from a noun/adjective.
- The quotative particle or the complementizer എന്ന് (*enn*) 'that' usually occurs as a suffix of the verb or the copula. Given that we split the copula as a syntactic word, we split the complementizer as well. (Also, it increases parallelism with languages where complementizers are independent words, and avoids having to define a language-specific feature for verb with complementizer.)
- The coordinating clitics -ഉം (*-uṁ*) 'and' and -ഓ (*-ō*) 'or' are written together with conjuncts but analyzed as separate syntactic words.
- In orthography sometimes the object and the verb of a sentence occur as a multiword token. For example, in the sentence പെൺകുട്ടി തന്റെ സുഹൃത്തിന് കത്തെഴുതി (*peṇkuṭṭi tanṟe suhrttin katteḻuti*) 'The girl wrote a letter to her friend', കത്ത് (*katt*) 'letter' and എഴുതി (*eḻuti*) 'wrote' occur as a multiword token and are split.

4.3. Annotation

The selected sentences from the IndicCorp were added to the CoNLL-U Editor. The editor commands were thereby used to carry out the annotations. Splitting of tokens and/or paragraphs[6] were done in the editor itself.

4.4. Validation and Feature Checking

The official UD validation script[7] verifies the CoNLL-U file format as well as data conformity with the general UD annotation guidelines. It can also check permitted feature-value combinations for individual part-of-speech categories in the given language, dependency relation subtypes, and lemmas of auxiliary verbs. We have provided Malayalam-specific definitions for these tests.

While the validator can exclude certain universally defined feature values from Malayalam data, and it can allow feature values separately for individual POS categories, we want to specify more detailed rules that go beyond this. For example, the UD validator knows that `Gender` is relevant for pronouns in Malayalam, but we want to make sure that it occurs only with third-person personal pronouns. The UD validator checks that `Tense` does not occur with anything but verbs (and auxiliaries), but we want to be more specific, allow it for indicative forms and disallow it for imperative and necessitative forms. Moreover, we want to increase consistency by requiring that all verbs in indicative have a non-empty value of `Tense`. Tests of this sort are implemented in the Udapi-Python tool[8] (Popel et al., 2017) in the processing block

[6]While normally a line in the corpus corresponds to one sentence, some lines were sequences of multiple sentences.

[7]`https://universaldependencies.org/release_checklist.html#validation`

[8]`http://udapi.github.io/`

`ud.ml.MarkFeatsBugs`. In the future we envisage similar language-specific tests also for the dependency relations.

4.5. UDPipe

Manual annotation is a laborious task, especially if all morphological features have to be filled out for every word in a morphologically rich language. We thus use UDPipe 1.2[9] (Straka and Straková, 2017), a trainable tool that can tokenize text, tag it and parse it in the UD style. Obviously, the output of UDPipe is not perfect, so we must invest significant manual effort anyway, but at least part of the annotation can be guessed correctly by UDPipe's model.

After annotating the first 30 sentences (which was done without the help of UDPipe), we used these sentences as training data and trained a simple model (with the default configuration). This model was then used to parse 100 sentences from IndicCorp. As expected, the accuracy was quite bad, but at least the tool could guess the approximate word segmentation and prepare the data in the CoNLL-U format. In the next round we carefully polished annotation of the new sentences until it passed the UD validation and all additional consistency tests defined by us. A new UDPipe model, trained on 130 hand-annotated sentences, was significantly better and could predict some annotations correctly. We will repeat this process with new batches of manually verified data and we expect the model to gradually improve and make fewer errors.

5. Part-of-Speech Tagging

The current version of the treebank contains 16 part-of-speech tags including `SYM` and `X` (see POS frequencies in Table 1); the only category missing from the current data is interjections. For the POS tagging the morphological cues were predominantly used. But in some cases the syntactic context was considered to capture the word category in a better way. For instance, the quotative particle എന്ന് (*enn*) 'that' is tagged `PART` where it is used as a 'quotative marker' and `SCONJ` where it is used as complementizer.

AUX: The copula verbs ആക് (*āk*) 'be' and ഉണ്ട് (*uṇṭ*) 'be' are tagged `AUX`. Additionally, the modal auxiliary verbs കഴിയുക (*kaḻiyuka*) 'can, be able to' and വേണം (*vēṇaṁ*) 'want' are also tagged `AUX`.

CCONJ: The particle -ഉം (*-uṁ*) 'and' that serves as a conjoining element for nouns and verbs is tagged `CCONJ` along with പക്ഷേ (*pakṣē*) 'but' and the particle -ഓ (*-ō*) 'or'. In Malayalam, the third person plural pronouns അവർ (*ivar*) 'they' and അവ (*iva*) 'these' can act as a conjunction if realized as എന്നിവർ (*ennivar*) and എന്നിവ (*enniva*). These forms are also tagged `CCONJ`.

[9] `https://ufal.mff.cuni.cz/udpipe/1`

POS	count	POS	count	POS	count	POS	count
ADJ	230	CCONJ	93	PART	58	SCONJ	25
ADP	38	DET	39	PRON	84	SYM	1
ADV	99	NOUN	720	PROPN	260	VERB	282
AUX	113	NUM	42	PUNCT	317	X	2

Table 1. Frequencies of POS tags.

SCONJ: In Malayalam, the reported speech is marked with a quotative particle (Asher and Kumari, 1997). The quotative particle എന്ന് (*enn*) when used as the complementizer is tagged `SCONJ`. Malayalam has only one sentence-final complementizer.

PART: The particle -ഉം (*-uṁ*) when used as an emphasizing element (rather than conjunction) is tagged `PART`. The quotative particle എന്ന് (*enn*) and its variant എന്ന (*enna*)[10] used in adnominal clauses are also tagged `PART`.

6. Morphological Features

The inherent gender of nouns[11] determines which personal pronoun can refer to the noun, and it is sometimes reflected as agreement on adjectives. It is not reflected on verbs (unlike in related Tamil). We do not annotate the gender of nouns in data but we do so for third-person pronouns with one of three values: `Masc`, `Fem` or `Neut`. Like `Gender`, `Animacy` is also an inherent feature of nominal words (`NOUN`, `PROPN`, and `PRON`). It has two values: `Anim` and `Inan`. Animacy is grammatically relevant because inanimate nouns may occur without accusative marking -എ (*-e*) when used as direct objects (cf. examples (1a) and (1b) below). Animates include nouns denoting persons and in some cases animals, or trees. Animacy aligns with gender only partially. Masculine and feminine third person pronouns refer to persons and are perceived as animate. Neuter pronouns can be animate if referring to animals or plants, and inanimate otherwise. For inanimates, the accusative form is equal to the nominative അത് (*at*) 'it', while for animates it uses a separate form അതിനെ (*atine*) 'it'. We annotate the animacy of third person neuter pronouns but we omit the feature for other personal pronouns.

In example (1) we can see how the accusative case assignment based on animacy of the objects plays a vital role in disambiguating the subject and the object. The example

[10]The quotative particle is realized as the relative particle എന്ന (*enna*) in relative clauses. It is referred to as 'relative particle' in Asher and Kumari (1997).

[11]There is a tendency that masculine nouns end in -അൻ (*-an*) and feminine nouns in -ഇ (*-i*). For example, male thief is കള്ളൻ (*kaḷḷan*) and female thief is കള്ളി (*kaḷḷi*). However this type of classification cannot be generalized (Asher and Kumari, 1997).

(1d) shows that if the object and subject both are animate, then the object needs to be marked accusative, otherwise it will not be possible to distinguish the subject and object in the sentence (because both SOV and OSV word orders are possible).

(1) a. ñān oru vaṇṭi vāṅṅi
I.NOM one car buy.PAST
'I bought a car'
b. ñān oru vaṇṭi(y)-e vāṅṅi
I.NOM one car-ACC buy.PAST
'I bought a car'
c. ñān avan-e viliccu
I.NOM he-ACC call.PAST
'I called him'
d. *ñān avan viliccu
I.NOM he call.PAST
'I called he'

Case has 13 possible values: `Nom`, `Acc`, `Gen`, `Dat`, `Ins`, `Loc`, `Abl`, `All`, `Cmp`, `Com`, `Ben`, `Cau`, `Voc`. Malayalam is an agglutinative language and many spatiotemporal and/or case-like morphemes are analyzed as postpositions. The `Case` feature occurs with the nominal words, i.e., `NOUN`, `PROPN`, `PRON`, `NUM` and also with nominalized verb forms. Nominalized verb forms are frequently used where the verbs take the nominalizing suffix -ത് (*-t*) (Asher and Kumari, 1997). These verb forms are marked as `VerbForm=Vnoun` and are morphologically marked for case, tense and polarity. In cleft constructions, they occur along with the copula ആക് (*āk*), which is postposed to the focused element. In example (2) we can see how nominalization works in Malayalam.

(2) a. avan at śariyāyi paṟaññu
he.NOM that correctly say.PAST
'He said it correctly'
b. avan paṟaññat śariy-āṇ
he.NOM say.PAST.NML correct-be.PRES
'What he said was correct.'
c. avan paṟaññat-āṇ śari
he.NOM say.PAST.NML-be.PRES correct
'*What he said* was correct.'

Example (2a) is a simple declarative clause with a finite verb. (2b) shows the nominalized construction and (2c) is a cleft construction.

7. Dependency Relations

The main dividing lines in the taxonomy of dependency relations in UD are between the core arguments of clausal predicates, non-core dependents of clausal predicates, and dependents of nominals.[12]

7.1. Core and Non-Core Dependents

According to the UD taxonomy, core arguments are subjects and objects. But this limits the treatment only to those constituents that are morphologically marked with the nominative and accusative case.[13] The non-core dependents or the oblique dependents are those arguments with coding strategies not used by the core arguments (Zeman, 2017). In world's languages, certain predicates would take dependents occupying the subject and object positions and not marked as nominative and accusative respectively. For example, in the Czech sentence *Martin hýbá nábytkem* 'Martin moves the furniture', the noun *nábytek* 'furniture' takes the instrumental case, although the verb *hýbat* 'to move' selects it as an argument. On being passivized the object remains in the instrumental case (Zeman, 2017). Similar examples from other languages show that what is traditionally regarded as 'objects' or 'subjects' in these languages may be coded with cases similar to the oblique dependents.

7.1.1. Non-Nominative Subjects

The constituent ordering in morphologically rich languages can be different from the *typical* ordering of nominative constituents preceding the non-nominative constituents and it is largely semantically predictable (Bayer, 2004). For example in German we do find instances where a dative argument occurs with certain predicates which may or may not have any nominative arguments.

(3) a. Mir ist kalt
me.DAT is cold
'I am cold'

b. Mir war schlecht
me.DAT was bad
'I was sick'

Data from Sigurðsson (2004) for Icelandic also shows similar constructions. In Icelandic the non-nominative subjects (NNS) are referred to as *quirky* subjects (Sigurðsson, 1992) as they pass the tests for subjecthood.

[12] https://universaldependencies.org/u/dep/

[13] This follows from the normal treatment of A and P arguments in primary transitive clauses (Andrews, 2007) in Malayalam. The nominative and accusative cases identify nominal arguments. For open and closed clausal dependents the core vs. non-core distinction is trickier (Przepiórkowski and Patejuk, 2018).

(4) a. Þeim er kalt
them.DAT is cold
'They are freezing'

b. Henni fór fram
her.DAT went forth
'She got better'

This type of pre-verbal dative arguments in German and Icelandic look similar, nevertheless they are syntactically different from each other (Fischer, 2004): In German they are just oblique dependents, while in Icelandic there is evidence that they behave like subjects, that is, core arguments.

Similar dative experiencer *subjects* in Kannada (Amritavalli, 2004) and Hindi (Mahajan, 2004) originate in unaccusative contexts, i.e., the nature of the predicates decides the origin of these non-nominative subjects. In Malayalam, the dative experiencer constructions occur with predicates that express possession and mental or physical experience.

(5) a. avalkk oru vīṭ uṇṭ
her.DAT one house is
'She has a house'

b. enikk viśakkunnu
me.DAT hunger.PRES
'I am hungry'

The dative case of the dative NPs in Malayalam is an inherent or a semantic case (Jayaseelan, 2004a) and there can be more than one case relation for an argument as in (6):

(6) a. enikk kaḻiy-illa, ninn-e nokk-ān
me.DAT be.able-NEG you-ACC look.after-INF
'I cannot look after you' (Jayaseelan, 2004a)

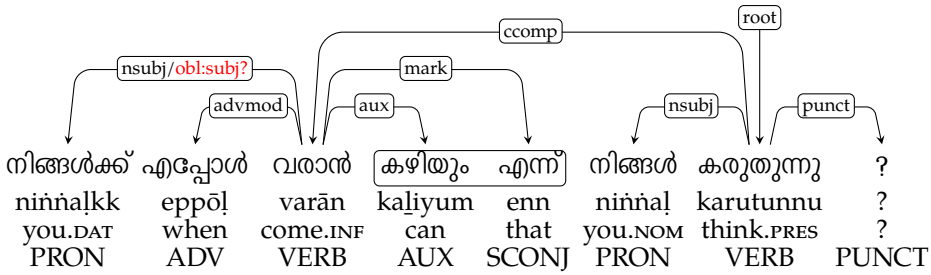
b. enn-ekkoṇṭu kaḻiy-illa, ninn-e nokk-ān
me.INSTR be.able-NEG you-ACC look.after-INF
'I cannot look after you' (Jayaseelan, 2004a)

With the verb നോക്കുക (*nōkkuka*) 'look after', we can have the dative and instrumental alternation on the *subject* argument. Both the sentences in (6) have the same semantic reading. With a different verb having different semantics this is not possible:

(7) a. enikk ninne iṣṭam alla
me.DAT you-ACC liking NEG
'I don't like you'
(Jayaseelan, 2004a)

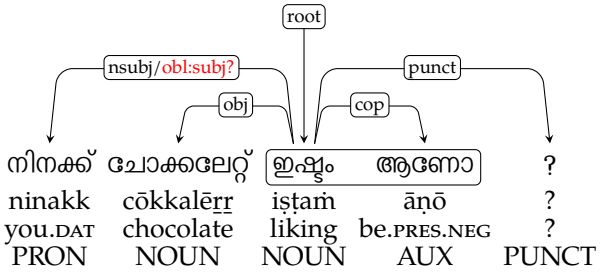
b. *enn-ekkoṇṭu ninne iṣṭam alla
me.INSTR you-ACC liking NEG
'I don't like you'
(Jayaseelan, 2004a)

Hence, Jayaseelan (2004a) concludes that dative NP is an oblique argument, not a subject as the case-marking of the verb's oblique arguments are semantically determined. Zeman (2017) has shown that the non-nominative arguments in Russian and Czech do not behave like *typical* subjects and should be treated as oblique arguments



'When do you think you can come?'

Figure 1. An example of a non-nominative subject.

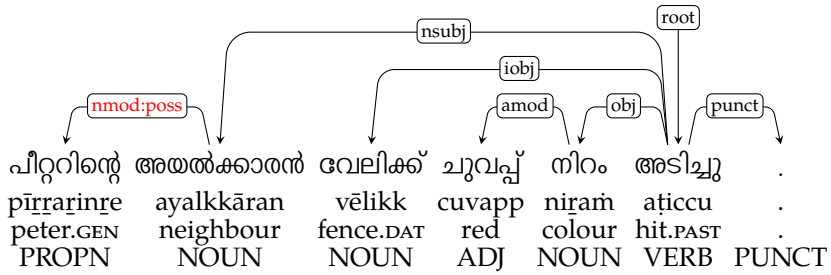


'Do you like chocolate?'

Figure 2. An example of a non-nominative subject.

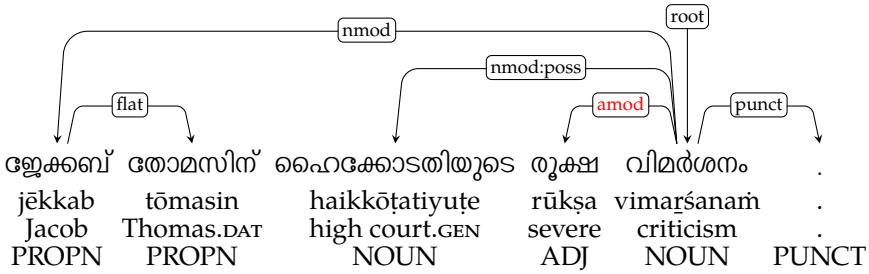
marked with the dependency relation `obl:arg`. However, the existing UD treebanks for the Dravidian languages, i.e., Tamil MWTT (Krishnamurthy and Sarveswaran, 2021) and Telugu MTG (Rama and Vajjala, 2018) treat the non-nominative arguments as core dependents marking them as `nsubj:nc`. We have tentatively also used the `nsubj` relation for non-nominative subjects in the current version of the Malayalam UD treebank, hence all Dravidian UD treebanks are compatible. However, we regard the question as open and do not exclude the possibility of re-analyzing them as oblique dependents in the future—preferably throughout the Dravidian family.

Example annotation of NNS in Malayalam is shown in Figures 1 and 2. Since the UD taxonomy of core vs. non-core dependents is an ongoing discussion in the UD community we may revert the NNS to oblique dependents and label them with a new subtype `obl:subj`. The goal here is to achieve a consistent explanation of the NNS constructions across the Dravidian languages.



'Peter's neighbor painted the fence red.'

Figure 3. An example of genitive modification.



'High Court criticizes Jacob Thomas.'

Figure 4. An example of an adjective modifying a nominal.

7.2. Nominal Dependents

nmod: We have used the `nmod` relation to mark the attributes of nouns or noun phrases. The label `nmod:poss` is used for the genitive complements (Figure 3).

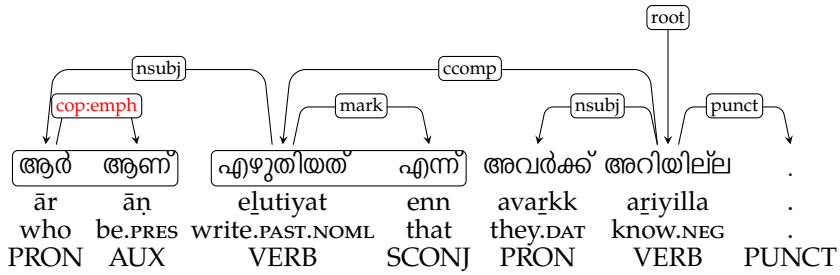
amod: This relation is used for all the non-clausal adjectival attributes of nouns or pronouns (Figure 4).

7.3. Other Dependency Relations

Here we discuss the other relations, mainly the subtypes[14] of various dependency relations that are used for Malayalam.

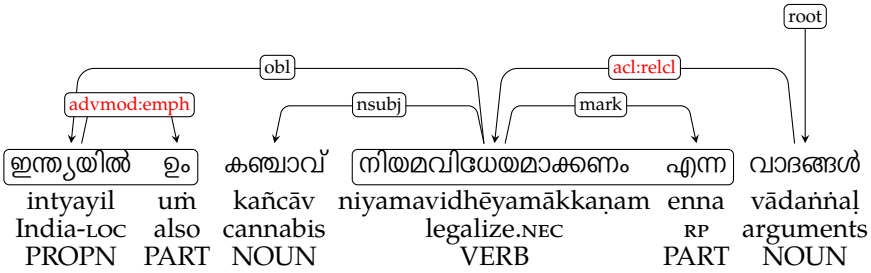
cop:emph is a special relation capturing the *focus* in a phrase. In cleft constructions, the verb is nominalized and the copula is postposed to the focused element (Figure 5).

[14]Subtypes are language-specific and optional.



'They do not know who wrote it.'

Figure 5. An example of a copula used to emphasize the focused constituent in cleft constructions. More literally, the sentence says 'Who is it (whose) writing (it was), that they know-not.'



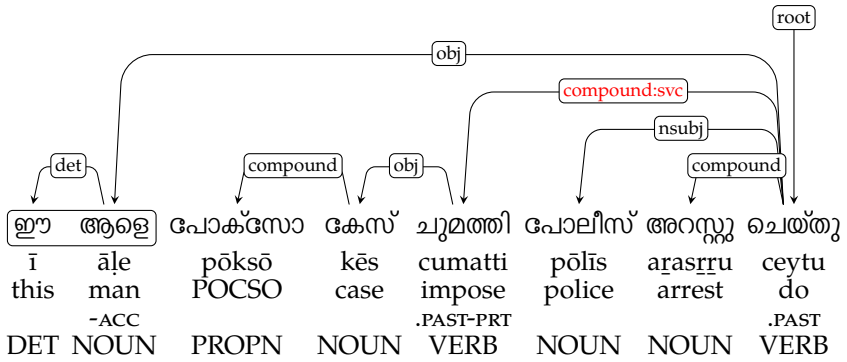
'Arguments to legalize cannabis in India'

Figure 6. An example of the -ഉം (uṁ) particle emphasizing a nominal. This sentence also serves as an example of a relative clause.

`advmod:emph`: The particle -ഉം (*uṁ*) (which is also the coordinating clitic) is used as an emphasizing element and to differentiate it from the `cc` dependency relation, `advmod:emph` is used (Figure 6).

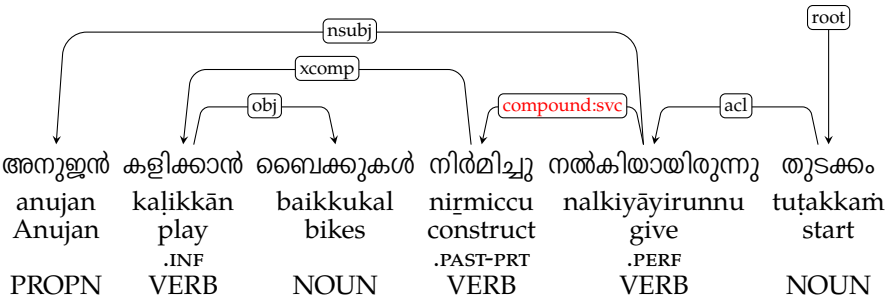
`compound:svc`: Serial verb constructions are the sequence of verbs and their (shared) complements.[15] The verbs in these constructions are not separated by any overt marker of coordination or subordination. In most of the cases, the verbs are lexicalized and cannot be separated by any intervening material. The final verb is usually finite and the preceding verbs are non-finite and resemble the past participle forms (Jayaseelan, 2004b) (Figures 7 and 8).

[15] `https://universaldependencies.org/u/dep/compound-svc.html`



'He was arrested by the police on a POCSO case.'

Figure 7. An example of a serial verb construction.

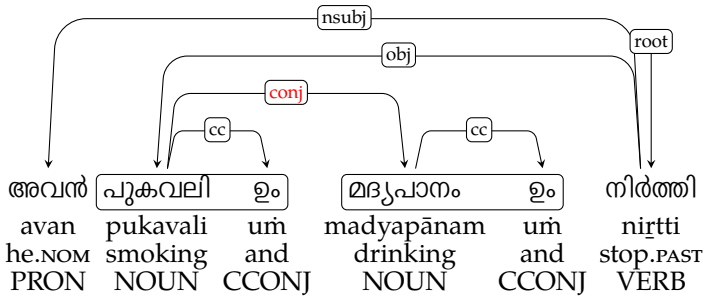


'It started with making bikes for Anujan to play with'

Figure 8. An example of a serial verb construction.

`acl:relcl`: The relative clause formation requires the relative particle എന്ന (*enna*). This subtype can be also applied to the sentential relative clauses but there are no such examples in the treebank yet. The participial relative clauses are treated as dependents of nominals and are labelled with the dependency relation `acl:relcl`; an example can be seen in Figure 6.

`conj`: According to UD guidelines, coordination receives asymmetric treatment, i.e., the leftmost conjunct is the technical head and all other conjuncts 'depend' on it. For head-final languages it may cause a problem as discussed for Japanese and Korean in Kanayama et al. (2018). This depends on how the case marking happens in these languages. If both the conjuncts are case-marked then the left-headed `conj`



'He stopped smoking and drinking'

Figure 9. An example of coordination.

relation works fine but if the mechanism of case assignment to the conjuncts happens in some other way that might disrupt the phrasal units, then the existing principle of left-headed `conj` may pose some challenges. In Malayalam, we see that the left-headedness does not cause any problems. Malayalam uses multiple `cc` relations in a coordination unit because the coordinating clitics -ഉം (*-uṁ*) 'and' and -ഓ (*-ō*) 'or' are appended to each of the conjuncts (Figure 9).

8. Conclusion

This paper presents the properties of a new UD-based treebank for Malayalam. We have discussed the annotation process along with elaborating on the various choices of the dependency relations. The UD treebanks of the Dravidian languages need to adopt a consistent annotation for syntactically similar constructions. We have illustrated various ways in which many syntactic phenomena in Malayalam have been tackled based on the existing UD guidelines. In the subsequent releases of the treebank, the annotations may undergo subtle improvements.

Acknowledgements

This work was supported by the grants 20-16819X (LUSyD) of the Czech Science Foundation; and LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic. In addition, the first author was supported by the Scholarship of the Ministry of Education, Youth and Sports in Support of Foreign Nationals' Study at Public Institutions of Higher Education in the Czech Republic (promulgated on 28 January 2014 under ref. No. MSMT-44726/2013).

Bibliography

Amritavalli, R. Experiencer datives in Kannada. In *Non-nominative Subjects: Volume 1*, pages 1–24. 2004. doi: 10.1075/tsl.60.03amr.

Andrews, Avery D. The Major Functions of the Noun Phrase. In Shopen, Timothy, editor, *Language Typology and Syntactic Description. Volume I: Clause Structure*, pages 132–223. Cambridge University Press, 2007. doi: 10.1017/CBO9780511619427.003.

Asher, R. E. and T. C. Kumari. *Malayalam*. Routledge Descriptive Grammars. Routledge, London, 1997. doi: 10.4324/9781315002217.

Bayer, Josef. Non-nominative subjects in comparison. In *Non-nominative Subjects: Volume 1*, page 49–76. 2004. doi: 10.1075/tsl.60.05bay.

de Marneffe, Marie-Catherine and Joakim Nivre. Dependency Grammar. *Annual Review of Linguistics*, 5(1):197–218, 2019. doi: 10.1146/annurev-linguistics-011718-011842. URL https://doi.org/10.1146/annurev-linguistics-011718-011842.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June 2021. doi: 10.1162/coli_a_00402. URL https://aclanthology.org/2021.cl-2.11.

Fischer, Susann. The diachronic relationship between quirky subjects and stylistic fronting. In *Non-nominative Subjects: Volume 1*, page 193–212. 2004. doi: 10.1075/tsl.60.11fis.

Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, and et al. Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0), 2020. URL http://hdl.handle.net/11234/1-3185. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Heinecke, Johannes. ConlluEditor: a fully graphical editor for Universal Dependencies treebank files. In *Universal Dependencies Workshop 2019*, Paris, 2019. doi: 10.18653/v1/W19-8010. URL https://github.com/Orange-OpenSource/conllueditor/.

Jayaseelan, K. The possessor — experiencer dative in Malayalam. In *Non-nominative Subjects: Volume 1*, page 227–244. 2004a. doi: 10.1017/CBO9781139003575.006.

Jayaseelan, K. The Serial Verb Construction in Malayalam. In *Clause Structure in South Asian Languages*, pages 67–91. Springer Netherlands, 01 2004b. ISBN 978-1-4020-2719-2. doi: 10.1007/978-1-4020-2719-2_3.

Jiang, Jingyang and Haitao Liu. Review of Lucien Tesnière, Elements of structural syntax, translated by Timothy Osborne and Sylvain Kahane, Amsterdam & Philadelphia, PA: John Benjamins, 2015. *Journal of Linguistics*, 51(3):705–709, 2015. ISSN 0022-2267. URL https://www.jstor.org/stable/26570750.

Kakkonen, Tuomo. Dependency treebanks: methods, annotation schemes and tools. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, pages 94–104, Joensuu, Finland, May 2006. University of Joensuu, Finland. URL https://aclanthology.org/W05-1714.

Kakwani, Divyanshu, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4948–4961, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL https://aclanthology.org/2020.findings-emnlp.445.

Kanayama, Hiroshi, Na-Rae Han, Masayuki Asahara, Jena Hwang, Yusuke Miyao, Jinho Choi, and Yuji Matsumoto. Coordinate Structures in Universal Dependencies for Head-final Languages. pages 75–84, 01 2018. doi: 10.18653/v1/W18-6009.

Krishnamurthy, Parameswari and Kengatharaiyer Sarveswaran. Towards Building a Modern Written Tamil Treebank. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 61–68, Sofia, Bulgaria, December 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.tlt-1.6.

Mahajan, Anoop K. On the origin of non-nominative subjects. In *Non-nominative Subjects: Volume 1*, page 283–299. 2004. doi: 10.1075/tsl.60.16mah.

Popel, Martin, Zdeněk Žabokrtský, and Martin Vojtek. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden, 2017.

Przepiórkowski, Adam and Agnieszka Patejuk. Arguments and Adjuncts in Universal Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3837–3852, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1324.

Rama, Taraka and Sowmya Vajjala. A Dependency Treebank for Telugu. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 119–128, Prague, Czechia, 2018. URL https://aclanthology.org/W17-7616.pdf.

Sigurðsson, Halldor Armann. The case of quirky subjects. Workingpaper, Department of Scandinavian Languages, Lund University, 1992.

Sigurðsson, Halldor Armann. Icelandic non-nominative subjects. In *Non-nominative Subjects: Volume 2*, page 137–159. 2004. doi: 10.1075/tsl.61.09sig.

Straka, Milan and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3009. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Zeman, Daniel. Core Arguments in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 287–296, Pisa, Italy, September 2017. Linköping University Electronic Press. URL https://aclanthology.org/W17-6532.

Address for correspondence:
Abishek Stephen
stephen@ufal.mff.cuni.cz
ÚFAL MFF UK
Malostranské náměstí 25, Praha, CZ-11800, Czechia

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 120 APRIL 2023 47–71

Transferring Word-Formation Networks Between Languages

Jonáš Vidra, Zdeněk Žabokrtský

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

We present a method for supervised cross-lingual construction of word-formation networks (WFNs). WFNs are resources capturing derivational, compositional and other relations between lexical units in a single language. Current state-of-the-art methods for automatically creating them typically rely on supervised or unsupervised pattern-matching of affixes in string representations of words, with few recent inroads into deep learning. All methods known to us work purely in a monolingual setting, limiting the use of higher-quality supervised models to high-resource languages. In this paper, we present two methods, one based on cross-lingual word alignments and translation and another based on cross-lingual word embeddings and neural networks. Both methods are capable of transfer of WFNs into languages for which no word-formational data are available. We evaluate our models on manually-annotated word-formation data from the Universal Derivations and UniMorph projects.

1. Introduction

A word-formation network is a dataset capturing information about how are lexemes created using derivation, compounding, conversion and other types of relations. Such networks can be created using various degrees of automatization. On one end of the spectrum, there are networks created by manually annotating the individual relations, resulting in a dataset that is highly precise, but either expensive to create or small in size.

In this article, we explore methods from the other, unsupervised, part of the scale: methods which do not require any human input or in-language annotations of word-formation relations. Instead, they transfer knowledge from existing word-formation networks in other languages. One method we present uses parallel texts and off-the-shelf tools for tokenization and lemmatization, another one uses cross-lingual word

 Corresponding author: vidra@ufal.mff.cuni.cz
Cite as: Jonáš Vidra, Zdeněk Žabokrtský. Transferring Word-Formation Networks Between Languages. The Prague Bulletin of Mathematical Linguistics No. 120, 2023, pp. 47–71. doi: 10.14712/00326585.027.

embeddings. Parallel texts are significantly more abundant and easier to obtain than word-formation annotations and they are available for more languages – compare the OPUS collection (Tiedemann, 2012), where just the OpenSubtitles corpus is available for 65 languages, to a survey of available word-formation networks listing only 63 resources for 22 languages (Kyjánek, 2018). Similarly, cross-lingual word embeddings can be created for dozens of languages, e.g. XLM-R (Conneau et al., 2020) is pretrained on 100 languages from the CommonCrawl dataset.

As a result, our methods should allow for a cheap and rapid creation of word-formation networks for many languages, although at a cost of lower quality. We hope that it is possible to emulate the successes of transfer learning methods used for other similar tasks in natural language processing, such as syntactic parsing (McDonald et al., 2011), part-of-speech tagging (Zhang et al., 2016) or lemmatization (Rosa and Žabokrtský, 2019).

The main idea behind our methods is that translation of text between languages is supposed to preserve the pragmatic meaning of texts and it usually preserves also the meaning of individual sentences and words. Since word-formational relations connect words with similar semantics and orthography, multiple possible target-language translations of a single source-language word are word-formationally related with a higher probability than randomly selected words. Moreover, many types of word-formational relations have parallels across languages. For example, actor nouns are typically derived from verbs – and if we take two such nouns from two languages, which are translations of one another, chances are that their predecessor verbs will also be translation equivalents (e.g. the Czech and English relations *opravit* ("to repair") → *opravář* ("repairman") are parallel, even though one uses derivation and the other one compounding). Therefore, we believe that some information about word-formation relations can be shared across languages.

In practice, the transferred networks are too small to be usable, but they can serve as synthetic training data for a supervised machine translation model, which extracts word-formation patterns found therein and finds more examples of them across a large lexicon, thereby improving the recall of the resulting network. Synthetic training data are widely used in deep learning, e.g. in machine translation (Sennrich et al., 2016; Zhang and Zong, 2016).

The pilot experiments presented in this paper focus on one-to-one relations between lexemes. We omit compounding altogether and simplify the task of creating a word-formation network to a task of assigning each lexeme a single *parent* lexeme, or deciding that it is unmotivated and should function as a root of the morphological family.

2. Related work

Most existing word-formation data is in the form of manually- or semi-automatically-created word-formation networks. These are made individually for each language,

using annotation schemas tailored towards that language's needs. Two larger projects aim to unify the annotation formats and provide data for more languages in a single format: Universal Derivations (Kyjánek et al., 2019) and, recently, UniMorph since version 4.0 (Batsuren et al., 2022).

Universal Derivations (UDer) extracts its data from word-formation networks created by linguists. The collection contains 31 resources covering 21 languages. Individual resources differ in annotation goals (some resources marking all word-formational relations, others e.g. only deverbal derivations), size (ranging from a thousand to a million lexemes), and quality. Some resources in the collection contain also other annotations, such as semantic labels of the relations or morphological segmentation.

UniMorph is a massively multilingual resource which aims at describing morphology in a general, language-universal way. The UniMorph data covers inflection of 168 languages, with 25 of them also containing word-formational information. The word-formational data, sourced from Wiktionary, describes derivational morphology only and contains no features other than derivational relations and annotation of the changed morpheme(s) in the successor lexeme. As with UDer, many datasets are small, covering only a few thousand relations.

In addition to the manually-created word-formation networks, multiple models for automatic construction have been proposed, typically working on the formal level (textual-string-wise) by detecting paradigmatic changes between the predecessor(s) and successor. Baranes and Sagot (2014) created a method that infers derivational relations from inflectional paradigms and reported a very high precision (80-98% depending on the language). The relations are detected by first extracting a list of possible prefixal and suffixal changes and then pattern-matching pairs of words against it. The inflectional paradigms are used for reducing problems with suppletion and allomorphy within stems, which would otherwise cause the prefix- and suffix pattern matching to fail – e.g. if we know that *spoken* is a past participle form of a lexeme with lemma *speak*, we can derive the lexeme *unspoken* from *speak* using the rule $X \rightarrow$ *un-*X.

A different solution to the problem of allomorphy is proposed by Lango et al. (2021), who use a pattern-mining method to detect rules of allomorphy jointly with affixation. The patterns are extracted automatically in an unsupervised fashion and the potential relations are ranked by a machine-learning model trained on a small manually annotated word-formation network.

Batsuren et al. (2019) deal with cognate detection (i.e. linking words of common origin, identical meaning and similar spelling in different languages) using a multilingual approach. The multilingual data they use is a specialized linguistic resource containing information about etymological ancestry, which means that their methods are not directly applicable in our semi-supervised setting.

Cognates can also be used as a clue for aligning parallel corpora and several methods for detecting cognate pairs were developed with the alignment task in mind, but these methods need not be very precise – e.g. Church (1993) uses identical character

4-grams and Simard et al. (1992) use pairs of words with identical first four characters; both methods are too imprecise to recognize exact word-formational relations.

More recently, algorithms working with word embeddings (as a proxy for a semantic representation) have also been proposed: Musil et al. (2019) show that word embedding differences between word-formationally related words reflect the word-formation paradigm of the relation, and perform clustering of word-formation relations to retrieve the paradigms, although they don't use the models to produce a word-formation network. Svoboda and Ševčíková (2022) use a fine-tuned Marian translation model (Junczys-Dowmunt et al., 2018) to directly produce parent lemma(s) for a given child lemma. The model requires a very large amount of data to train, and they solve this issue by creating synthetic training data with a simple manually-crafted morphology model, which creates nonsensical, but well-formed compounds. This works, because the focus of PareNT is Czech compounding, which has a simple formal structure, unlike typical derivational patterns in most languages.

The task of constructing word-formation networks is superficially similar to the task of dependency parsing – in both cases, one tries to attach words to typically a single parent (head or predecessor). However, there are also important differences: Dependency parsing is in many ways computationally simpler, because the space of potential heads for any single lexical unit is bounded by the length of a sentence (typically tens of units), while in WFN construction, any lexeme in the language can be the correct predecessor (typically hundreds of thousands of units). Also, when machine learning is used, data for syntactic parsing is more abundant, because the inventory of training sentences is potentially infinite and getting new ones from a corpus is relatively cheap, while with WFNs, the number of training examples is limited to the lexicon size.

3. Models

Our models process data in two steps: In the first step (projection), a cross-lingual method is used to create a small word-formation network in the target language using training data in other languages. In the second step (bootstrap extension), the small network from the first step is used as synthetic training data to train a supervised model of word formation, which produces a large word-formation network.

In this paper, we present two models for each step: The projection step can be performed either by the Transfer model (see 3.1.1), or by the Cross-lingual embedding model (see 3.1.2). The bootstrap extension step can be performed by the Statistical machine learning extension model (see 3.2.1) or by the Neural extension model (see 3.2.2).

3.1. Projection models

3.1.1. Transfer model

To transfer a word-formation network from a source to a target language, we view the network as a list of parent-child derivational relations and attempt to find the best parent for each target-side lexeme using a word-translation model together with target-side formal similarity metrics. Conceptually, the input lexeme C is first back-translated into the source language as C', a suitable parent P' of the translation is found in the source word-formation network and this parent is translated into the target language as P.

The translations and backtranslations are found using a probabilistic word translation lexicon induced from word-aligned data obtained by running FastAlign (Dyer et al., 2013) on a lemmatized parallel corpus. Since the present article does not consider compounding, univerbation or other word-formation relations connecting more than two lexemes, we count each pair of aligned lexemes separately, regardless of whether one of the lexemes has other alignments in that parallel sentence pair. As a result, a lexeme aligned to a multi-word phrase is considered to be equally translated from each member lexeme of that phrase.

Since there may be multiple possible translations of each lexeme, and because the most suitable parent needn't be the direct parent of C', but rather another member of its word-formational family (e.g. the Czech lexemes *svoboda* ("freedom") → *svobodný* ("free") have the opposite derivational relation from English or German *frei* → *die Freiheit*), the process is conducted probabilistically, yielding many potential parents P for each C, each with a score. The target network is then found by finding the spanning tree of this graph of relations which maximizes the product of the scores (Chu and Liu, 1965).

The score s of each potential relation $P \rightarrow C$ is obtained as a weighted arithmetic mean (with weight w) of the translation score $\mathrm{Xfer}(C, P)$ and a relative edit distance computed from the Levenshtein distance $l(C, P)$, according to Equation 1 below. The relative edit distance is the Levenshtein distance between the lemmas of C and P divided by the maximum of their lengths, yielding a number between 0 and 1.

$$s = \frac{\mathrm{Xfer}(C, P) + w \cdot (1 - \frac{l(C,P)}{\max(|C|,|P|)})}{w + 1} \tag{1}$$

We define the translation score of C and P as $\mathrm{Xfer}(C, P)$ according to Equation 2, where $|\mathrm{align}(x, y)|$ denotes the number of alignments between lexemes x and y seen in the aligned data and $\mathrm{dist}(C', P')$ denotes the number of relations on the shortest path from C' to P' in the source network.

$$\mathrm{Xfer}(C, P) = \sum_{\forall C', P'} \frac{|\mathrm{align}(C, C')|}{\sum_{\forall x} |\mathrm{align}(C, x)|} \cdot 0.5^{\mathrm{dist}(C', P')} \cdot \frac{|\mathrm{align}(P', P)|}{\sum_{\forall x} |\mathrm{align}(P', x)|} \tag{2}$$

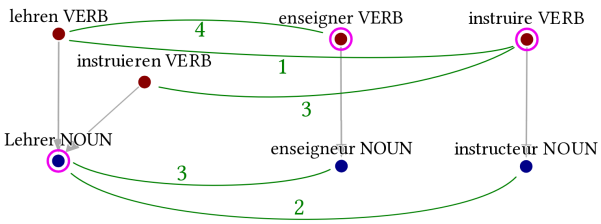


Figure 1: An example of finding a parent for the German lexeme *Lehrer* ("teacher") by transferring information from a French word-formation network, with word-formation relations in grey and alignments in green. *Lehrer* is aligned to *enseigneur* $^{3}/_{5}$ times, which has *enseigner* available through 1 relation, to which *lehren* is aligned $^{4}/_{4}$ times. *Lehrer* is also aligned to *instructeur* $^{2}/_{5}$ times, which has *instruire* available through 1 relation, to which *lehren* is aligned $^{1}/_{4}$ times and *instruieren* $^{3}/_{4}$ times. The translation score of *lehren* $\rightarrow$ *Lehrer*, calculated according to Equation 2 below, is therefore $\frac{3}{5} \cdot \frac{1}{2} \cdot \frac{4}{4} + \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{4} = 0.35$ while the score of *instruieren* $\rightarrow$ *Lehrer* is $\frac{2}{5} \cdot \frac{1}{2} \cdot \frac{3}{4} = 0.15$. The relative edit distance is $^{2}/_{6}$ for *lehren* $\rightarrow$ *Lehrer*, and $^{8}/_{11}$ for *instruieren* $\rightarrow$ *Lehrer*. Therefore, the final score of *lehren* $\rightarrow$ *Lehrer* is $\frac{0.35+5\cdot(1-2/6)}{6} = 0.336$ and the score of *instruieren* $\rightarrow$ *Lehrer* is $\frac{0.15+5\cdot(1-8/11)}{6} = 0.252$.

Therefore, the translation score is the product of the conditional probability of obtaining the backtranslated lexeme C′ given the lexeme C and the conditional probability of obtaining the translated parent lexeme P given P′, halved for each relation that has to be traversed between C′ and P′. If there are multiple possible choices of C′ and P′ for the given C and P, their translation scores are summed.

To prevent relations with low scores from being selected in the case where there are no better candidates, a relation is only considered for inclusion if its score is higher than a threshold.

An illustration of the translation score calculation is given in Figure 1.

The transfer algorithm is parametrized by the weights used for calculating the weighted mean of the translation and edit distance scores, and by the threshold. Since we intend to use the transfer algorithm in an unsupervised setting, it is necessary to obtain the weights without training them using e.g. grid search or gradient descent on in-language annotations. We have, however, found that although the algorithm is moderately sensitive to the setting of the weights and the threshold, the optimal settings in all tested languages are nearly identical. This allows us to train the hyperparameters on one language pair in a supervised manner and use them on other pairs without further training. Using grid search on the Czech $\rightarrow$ German transfer pair, we set the weight of the edit distance to 5, the weight of the translation to 1 and the threshold to 0.8.

3.1.2. Cross-lingual embedding model

Our second proposed model is a pairwise classifier neural network. Its inputs are two lexemes (l_p the potential predecessor, l_s the potential successor) represented by their word and character embeddings, and the output is a score classifying l_s as derived from l_p or not[1].

The model uses non-pretrained character embeddings, which are merged to produce word-level states by passing them through a bidirectional recurrent neural network layer with GRU activation. The resulting GRU states are concatenated with pretrained word embeddings and passed through a single hidden layer with ReLU activation. The hidden states are then classified using SoftMax into two classes, derived or nonderived.

The architecture can be used both monolingually and cross-lingually, if cross-lingual word embeddings are available. In our case, cross-lingual training is used to obtain synthetic training data for each language, followed by either a second cross-lingual phase utilizing all synthetic datasets to train a single model, or a monolingual training phase training one model for each language. In the first cross-lingual phase, the model for each target language is trained separately, using only data from other languages. No model is therefore trained on the same language it predicts data for, simulating a semi-low-resource setting in which raw texts are available for training the word embeddings, but annotated word-formation data is missing.

Since the network classifies pairs of lexemes and classifying all pairs in the lexicon is computationally prohibitively expensive (the complexity increases quadratically with the lexicon size and the larger networks have on the order of 1 000 000 lexemes), the following heuristic is used to preselect pairs with a long common prefix and suffix. The lexicon is alphabetically sorted in a prograde and retrograde fashion, and for each lexeme, we test potential predecessors that lie within a 10 lexeme window around it in either sorting, for a total of 40 potential relations.

The selected lexemes needn't be the 40 ones with the longest common prefixes and suffixes, but at least 10 are guaranteed to share the longest prefix and 10 the longest suffix. We perform the lexicographic sorting on uppercased lemmas stripped of accent marks so that e.g. the German word *Wunsch* ("a wish") sorts close to *wünschen* ("to wish") despite the differences in case and the presence or absence of the umlaut.

This method of obtaining relation candidates depends on the linguistic properties of the languages under consideration. It works well with languages which derive words predominantly by affixation, with limited allomorphy in the stem and only rare circumfixation, apophony or suppletive relations, which this method generally doesn't detect as possible relations. Therefore, the preselection of classification exam-

[1]Two alternative formulations were considered and tested: A network directly producing l_p from a given l_s as a string of characters, and a network classifying a bag of potential successor words at once from a given l_p, but the model detailed above was found to outperform them from the start and research into the alternative architectures was not pursued further.

ples limits the potential performance of the models – if the true word-formational predecessor doesn't lie in the window of tested examples, it cannot be correctly classified by the model. However, testing has shown that the languages we selected for evaluation (see Section 4) all get reasonable results despite the simplicity of the method – in all gold-standard networks, the heuristic selects at least 90 % of true predecessors. Therefore, the relatively small window size is not the main limiting factor of the prediction performance.

For example, looking at a window of ± 5 lexemes catches 85 % of all possible derivational relations in the German DErivBase word-formation network and ± 10 catches 90 %. In the French Démonette network, 96 % of derivations are within ± 5 and 98 % are within a ± 10 window. In Czech DeriNet, a window of ± 5 contains 85 % of all relations and ± 10 contains 90 %. The method would perform poorly on languages with more frequent circumfixation or nonconcatenative morphology, such as transfixation or templatic morphology found in e.g. Hebrew or Arabic.

A possible systematic fix for detecting words derived by circumfixation would be to use a more complex measure of morphological similarity. A method we tried is the orthographic part of the model from Proxinette (Hathout, 2008), which approximates morphological relatedness by counting common n-grams of varying length, probabilistically weighting them by rarity in the corpus. Its construction allows enumerating lexemes most similar to an input lexeme in a computationally-tractable way, without considering all pairs. However, it produced inferior results on the Czech, German and French datasets we evaluated it on, and therefore we don't use it in our experiments.

A word-formation network is then constructed by calculating the maximum spanning tree of edges with a classification score ≥ 0.5, with the score used as the edge weight.

When training the network, the training data is sampled uniformly randomly from all positive examples in the training word-formation networks, supplemented by two sources of negative examples: Non-predecessor lexemes randomly sampled from the whole lexicon and non-predecessors sampled from the heuristic window around each lexeme. For each positive example pair, we sample 2 random negative pairs from the whole lexicon, and 3 random lexemes and for each of them one random non-predecessor from the window, for a 1:5 positive:negative sample ratio.

The word embeddings used are based on multilingually-aligned staticized XLM-R (Hämmerl et al., 2022). The XLM-R model (Conneau et al., 2020) provides high quality cross-lingual contextual embeddings, but since the task of identifying word-formational relations is lexical in nature, it is better suited for use with static embeddings. These are obtained using the X2Static method (Gupta and Jaggi, 2021), which distills static embeddings from contextual by a process similar to FastText's "continuous bag of words" (Bojanowski et al., 2017), but applied to contextual word embeddings instead of the words themselves. The staticization process transforms the

embeddings one language at a time, so the cross-lingual relations of embeddings are partially lost, requiring realignment using VecMap.

3.2. Bootstrap extension

One issue with the aforementioned methods is that the word-formation networks they are able to produce are limited in size, because they both work only on lexemes with large-enough frequency in a corpus. Therefore, it is desirable to increase coverage of lower-frequency parts of the lexicon and lexemes not seen in the parallel data or embeddings lexicon. We propose two different methods to do this, both trained on data produced by one of the methods above. The first method is based on statistical machine learning with manually selected features, the second one reuses the neural network described above in a different setting.

3.2.1. Statistical machine learning extension model

One way of increasing recall of the produced word-formation networks is to take the networks created by the transfer model or cross-lingual embedding model described above, extract affixal patterns found therein and apply them to a larger lexicon.

The affixal pattern of a (proposed) word-formational relation is an unsupervised approximation of the morpheme difference between the related lexemes. We obtain it as the leftover substrings to the left and right of the longest common contiguous substring shared by lowercased lemmas of the lexemes. For example, the relation *Kampf* ("a fight") → *kämpfen* ("to fight") has the longest common contiguous substring *mpf* and affixal pattern *ka-* → *kä-* + *-en*.

We use the transferred network as a seed to train a machine learning method to predict derivational relations by classifying pairs of lexemes as either directly derived or non-derived from one another. The output network is obtained by finding the maximum spanning tree of the graph of predictions (Chu and Liu, 1965). The features used for classification are the one-hot-encoded part-of-speech categories of both lexemes, their edit distance, the difference of their lengths, whether each of them starts with a capital letter and the frequency of their affixal pattern as seen in the training dataset.

Since this method works by classifying pairs of lexemes, we again use the heuristic method for preselecting classification pairs described in Section 3.1.2 to decrease the computational complexity.

We evaluated multiple classification methods implemented in the scikit-learn package (Pedregosa et al., 2011), namely SVC, LogisticRegression, AdaBoostClassifier, KNeighborsClassifier, DecisionTreeClassifier, BernoulliNB and Perceptron and selected logistic regression for consistent evaluation performance.

3.2.2. Neural extension model

The neural extension model reuses the architecture of the cross-lingual embedding model, but with different data. It is trained on synthetic word-formation networks produced by the cross-lingual embedding model described above. The use of the training data is different too.

The cross-lingual embedding model doesn't train on data for the language it predicts relations for to ensure correct separation between training and evaluation data, and therefore n models are trained to produce data for n languages. The extension model is fully cross-lingual – a single universal model is trained jointly on all languages and can classify word-formational relations for any language.

In addition to getting the benefit of supervised training on the target language, the neural extension model also benefits from an extended lexicon compared to the cross-lingual embedding model – while the dataset for the cross-lingual embedding model contains the intersection of the manually created WFNs with the embedding lexicon, the extension model uses the embedding lexicons directly, providing potentially more training examples.

4. Training and evaluation data

For training and evaluating the word-formation models, we use word-formational data from the Universal Derivations (Kyjánek et al., 2019) and UniMorph (Batsuren et al., 2022) projects.

The word embedding data required by the neural models is taken from pretrained X2S-MA (Hämmerl et al., 2022), which is a static embedding resource created from XLM-R (Conneau et al., 2020) by first staticizing the embeddings using X2Static (Gupta and Jaggi, 2021) and then realigning the resulting static embeddings cross-lingually. Although it doesn't use subword segmentation, and is limited to its training lexicon as a result, we've found it to outperform other sources of embeddings.

All resources mentioned above are available for many languages: UDer for 21, UniMorph for 25 and X2S-MA for 40. However, their intersection is more limited – only 14 languages have both a word-formation network (from either UDer or UniMorph) and pretrained embeddings available. From those languages, we selected the 13 listed in Table 1 for use with the neural-networks-based models. One language, Dutch, was excluded, because its word-formation network as contained in UDer has quality too low to be usable for either training or evaluation due to errors introduced in the UDer conversion process.

When there are multiple networks for one language, we train on concatenation of lists of all relations. Compounding relations are treated as multiple derivational relations with the same successor. The data sizes of the individual word-formation resources for languages which are also present in the X2S-MA embeddings dataset are summarized in Table 1.

Lang	Resource	Lexemes	Relations
deu	DErivBase*	280775	43367
deu	UniMorph	40155	29381
eng	CatVar*	82675	24628
eng	UniMorph	264690	225131
eng	WordNet	13813	7855
est	EstWordNet*	988	507
fas	DeriNetFA*	43357	35745
fin	FinnWordNet*	20035	11890
fin	UniMorph	48499	36997
fra	Demonette*	22060	13808
fra	UniMorph	93382	73259
hun	UniMorph*	38441	32477
ita	DerIvaTario*	8267	1783
kaz	UniMorph*	3158	1965
por	EtymWordNetPT	2797	1610
por	NomLexPT*	7020	4201
por	UniMorph	19236	12687
rus	DeriNetRU	337632	164725
rus	DerivBaseRU*	270473	134024
rus	EtymWordNetRU	4005	3227
rus	GCompAna	4931	1639
rus	UniMorph	19823	14048
spa	DeriNetES*	151173	42825
spa	UniMorph	42760	31293
tur	EtymWordNetTR*	7775	5838
tur	UniMorph	2836	1776

Table 1: Data sizes of different resources. Resources labelled UniMorph are Wiktionary data extracted by the UniMorph project (Batsuren et al., 2022), all other resources are taken from the UDer project (Kyjánek et al., 2019). Resources marked by a star are used for evaluation in addition to training.

The gold standard data for each language is always taken from one resource, even when multiple resources for the language exist, to avoid having multiple conflicting golden predecessors for a single lexeme. The datasets designated as golden are marked in Table 1 by a star. Due to the setup of the experiments, the resource used for evaluation on a language is never directly used for training of that particular language's model. However, in the second multilingual step, the resource is trained on indirectly, because models for other languages do use it. For example, the Portuguese

data are left out when training the Portuguese model, but are used for training the English model. The second level model then uses both English and Portuguese data from the previous step. We deem that this is not an issue, because the neural network cannot get high scores by reproducing its training data, as the data are transferred cross-lingually twice before evaluation.

The transfer model was trained and evaluated on three languages only, namely Czech, French and German. These languages were selected for the large size and quality of their word-formation networks as present in UDer – DeriNet 2.0 (Žabokrtský et al., 2016) with 809 282 relations, Démonette 1.2 (Hathout and Namer, 2014) with 13 808 relations and DErivBase 2.0 (Zeller et al., 2013) with 43 368 relations, respectively. The transfer model fails to extract useful information from source data with low accuracy, and since (unlike the neural model) it works purely on pairs of languages, it is not possible to combine smaller resources for multiple languages to get one larger usable dataset.

We transferred each network into both other languages and compared the result to the existing network for that language. The transfer was realized using word dictionaries obtained from word alignments of parallel data. We used the OpenSubtitles dataset from the OPUS collection (Tiedemann, 2012) for all language pairs, lemmatizing them with UDPipe 1.2 (Straka and Straková, 2017) and extracting only words tagged as adjectives, adverbs, nouns and verbs. The lemmatizer uses pretrained models trained on treebanks from Universal Dependencies (Nivre et al., 2016). The lemmatized corpora are then aligned using FastAlign (Dyer et al., 2013). The data sizes are listed in Table 2.

Lang pair	Sentences	Tokens on left	Tokens on right
de — cs	15 237 340	48 320 109	45 922 280
fr — cs	25 838 124	83 108 504	87 983 667
fr — de	14 779 572	44 135 610	48 440 995

Table 2: Sizes of parallel data for each language pair after part-of-speech category filtering.

5. Evaluation Method

We evaluate the performance of our systems by measuring precision, recall and accuracy in the task of assigning a parent to a lexeme. We define precision as the ratio of correctly predicted relations to all predicted relations, recall as the ratio of correctly predicted relations to all gold relations and accuracy as the ratio of correctly assigned parents or correctly recognized unmotivated lexemes to all gold lexemes.

```
for gold_child in gold.lexemes:
  if not gold_child.parent:
    true_negative++
  else:
    for t_child in translations(gold_child):
      for t_parent in family(t_child):
        for parent in backtranslations(t_parent, gold_child):
          if parent = gold_child.parent:
            true_positive++
            continue_line 1
    false_negative++
  accuracy := ((true_positive + true_negative) / (true_positive +
    ↪ true_negative + false_negative))
  recall := true_positive / (true_positive + false_negative)
```

Listing 1: Pseudocode for calculating oracle accuracy and recall of the transfer algorithm. The backtranslation function returns all backtranslations of `t_parent`, except those that translate to `gold_child`.

Therefore, the precision and recall don't take into account unmotivated lexemes, while the accuracy does. The gold-standard data is taken from the existing word-formation network for the target language.

Because the set of lexemes captured in the cross-lingually projected network differs from the one used in the gold-standard data, we calculate the metrics in two ways, which differ in their treatment of missing lexemes. "External" measures consider all gold-standard relations of lexemes missing from the evaluated network to be false negatives, while the "internal" measures ignore them instead measures and only measure scores on the intersection of the two lexicons. Therefore, the external measures quantify how close the method gets to reproducing the gold-standard data, while the internal scores show how good is the output itself. Precision is the same for both methods, but recall and accuracy differ. The baseline measures and the networks obtained by machine learning are created from the set of lexemes found in the gold-standard network, which makes the internal and external measures identical.

5.1. Baselines

To establish a lower bound of reasonably achievable scores, we created two baselines: one trivial, called "empty", and one inspired by the purely left- or right-branching parse, the standard baseline in syntactic parsing, called "closest-shorter".

The empty baseline for a given lexicon is calculated as the scores of an empty word-formation network created over that lexicon, i.e. a network without any relations. The

lexemes from gold-standard data which have no assigned parent are therefore evaluated as correct, while all lexemes with parents are incorrect, resulting in unmeasurable (zero) precision, zero recall and moderate-to-high accuracy.

The closest-shorter baseline gives each lexeme four options for its parent and selects the one which has a shorter lemma and the closest orthographic distance, as measured by the ratio of the length of the longest common contiguous substring to the sum of lengths of the two lemmas. The options to choose from are the previous and next lexemes in prograde sorting of the lexicon, and the previous and next lexemes in retrograde sorting. The lemma length criterion means that lexemes surrounded by longer neighbors in both prograde and retrograde sorting of the lexicon remain unmotivated. We have already observed that both ends of most derivational relations lie within a small window on a sorted lexicon, making this baseline rather strong in terms of both precision and recall.

5.2. Oracle Score

As an additional measure of the potential quality of the transfer approach, we measured the oracle score of obtaining the gold-standard parent through any combination of back- and forward-translations of gold-standard child lexemes. Under this measure, unmotivated lexemes are always considered to be correct, and a derived lexeme is considered to be correctly connected to its parent if it can be backtranslated to a member of a word-formational family, which contains a member that can be translated to the correct parent. The pseudocode of this algorithm is present in Listing 1. The recall and accuracy obtained using this algorithm represent the maximum scores achievable with the transfer method, if it selected the gold parent for each lexeme every time it is available.

Any error in the recall can be broken down into three categories: first, where we cannot translate the child to the language of the transferring network; (no `t_child` on line 5 of Listing 1); second, where there are no translations of any members of the translated lexeme's family (no `parent` on line 7) and third, where no possible parent matches the gold one (predicate on line 8 is always false).

6. Evaluation Results

As can be seen in Table 3, the networks created by the transfer algorithm are rather small in size. Within the constructed network, precision and recall are moderate for most language pairs, but when compared to the gold standard data, recall is nearly zero for all of them.

To a large degree, difference in scores between languages can be attributed to the testing data – each language has its own independently developed dataset with different design decisions, size and quality. Even datasets with identical names (DerivBase, DeriNet) were typically created by different teams working with different constraints.

		Size [k]		Internal scores [%]				Gold scores [%]		
Alg	Lang pair	Lex	Rel	Prec.	Recall	F1	Acc.	Recall	F1	Acc.
Xfer	de → cs	18	6.0	40	33	36	54	0.29	0.58	1.2
	fr → cs	20	7.0	42	36	39	54	0.37	0.73	1.3
	cs → de	14	3.8	27	35	31	66	2.5	4.5	18
	fr → de	3	0.6	14	14	14	65	0.20	0.39	4.2
	cs → fr	3	1.2	24	31	27	43	2.1	3.9	7.7
	de → fr	0.4	0.1	3.5	11	5.3	59	0.04	0.07	1.8
ML	de → cs	1 026	743	46	74	56	49	74	56	49
	fr → cs	1 026	743	40	70	51	44	70	51	44
	cs → de	280	68	35	68	46	80	68	46	80
	fr → de	280	35	44	39	42	85	39	42	85
	cs → fr	21	15	60	89	72	66	89	72	66
	de → fr	21	5	36	14	20	37	14	20	37
closest-shorter baseline	cs	1 026	809	21	54	30	23	54	30	23
	de	280	225	5.2	57	10	21	57	10	21
	fr	21	17	32	83	46	39	83	43	39
empty baseline	cs	1 026	0	N/A	0.00	0.00	21	0.00	0.00	21
	de	280	0	N/A	0.00	0.00	85	0.00	0.00	85
	fr	21	0	N/A	0.00	0.00	35	0.00	0.00	35

Table 3: Evaluation scores of the results and baselines for each language pair. The lexeme and relation counts are in thousands. Internal scores are measured on the set of lexemes in the generated network, gold scores on the set of lexemes from gold data. Precision is identical for both. For the machine learning and baseline algorithms, the distinction between internal and gold scores does not matter, since the lexicon used for prediction is taken from the gold-standard data as is.

	Scores [%]		Error cause [%]			WFN rel count	
Lang pair	Recall	Acc.	No child trans	No parent trans	No match	Xferred	Gold
de → cs	5.1	29	91	0.08	3.8	43 368	809 282
fr → cs	6.8	32	90	0.05	3.6	13 808	809 282
cs → de	34	90	52	0.23	13	809 282	43 368
fr → de	26	93	51	0.02	22	13 808	43 368
cs → fr	35	80	57	0.20	8.3	809 282	13 808
de → fr	22	64	62	0.07	16	43 368	13 808

Table 4: Transfer oracle scores for each language pair. Precision is 100% in all cases. The error causes list percentage of cases where the lexeme cannot be translated to the language of the transferring network, where no possible parents can be translated back, and when none of the translated parents match the gold one, respectively. The error percentage points are relative to the total relation count, i.e. they sum up to 100 together with recall. The last two columns list sizes of the transferred and gold-standard word-formation networks, measured in relations.

For example, whether unconnected lexemes remain in the database or are elided has a dominating effect on accuracy – accuracy on the German, English and Spanish gold-standard WFNs is higher than on the other ones, because they contain > 70 % lexemes without parents, which are comparatively easy to correctly predict, but don't contribute to either precision or recall.

The classification example preselection heuristic may be a bottleneck on performance, as it limits recall to approximately 90 % and several networks come rather close to that number. But it is still possible to improve performance by a large margin before being strictly limited by the heuristic.

The performance of the transfer method depends a lot on the size of the transferred network. Since the Czech DeriNet is an order of magnitude larger than the other networks, the gold scores for networks created by using it as a base are the highest ones, but even these don't match more than 2.5% of relations from the gold-standard data.

The precision of the constructed networks is also influenced by the translation quality. The alignment data trained on the deu-fra pair (in both directions) has many incorrect alignments. This doesn't affect the oracle score, since the correct translations will generally be found, but the wide distribution of the probability mass hurts the actual algorithm, which is unable to distinguish plausible and implausible translations.

		Internal scores [%]				Gold scores [%]		
Lang	Size	Prec.	Rec.	F1	Acc.	Rec.	F1	Acc.
deu	15765	14	34	20	42	5.7	8.1	20
eng	12819	30	47	37	47	18	22	30
fas	13877	22	82	35	31	11	15	13
fin	1616	16	10	13	38	2.4	4.2	15
fra	3767	27	66	39	35	7.9	12	11
hun	6158	41	35	37	23	8.7	14	7.7
ita	3465	9.3	57	16	27	24	13	23
kaz	522	58	37	46	30	17	27	16
por	2428	37	43	40	36	25	30	27
rus	8708	12	16	14	28	0.8	1.5	3.6
spa	13702	24	76	37	42	8.3	12	15
tur	2829	28	75	40	38	17	21	19

Table 5: Evaluation scores of the synthetic training data. Accuracy, precision and recall are in percent, size indicates the number of predicted relations.

		Internal scores [%]				Gold scores [%]		
Lang	Size	Prec.	Rec.	F1	Acc.	Rec.	F1	Acc.
deu	15	20	0.03	0.06	67	0.01	0.01	33
eng	6843	40	28	33	55	12	18	36
fas	186	15	0.08	0.51	35	0.08	0.16	14
fin	3864	16	40	23	28	6.1	8.8	11
fra	4241	24	80	37	32	8.1	12	9.7
hun	4765	54	30	38	24	8.5	15	8.0
ita	3816	9.3	72	16	23	28	14	20
kaz	830	57	70	63	46	30	39	24
por	1908	49	42	45	47	25	33	35
rus	11508	14	29	19	25	1.3	2.3	3.3
spa	13961	23	77	36	41	8.1	12	15
tur	2961	30	80	44	38	19	23	20

Table 6: Evaluation scores of the neural extension model applied on word-formation networks obtained by the cross-lingual embedding model. Accuracy, precision and recall are in percent, size indicates the number of predicted relations.

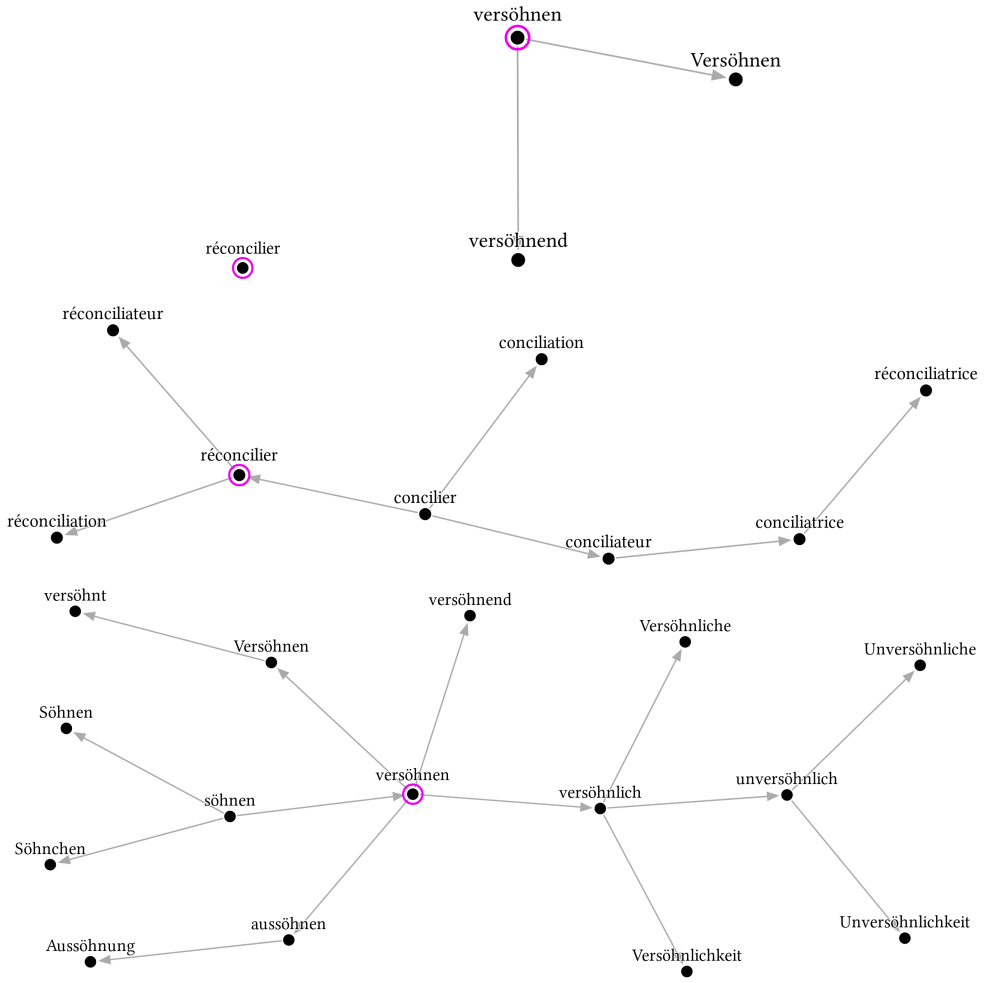


Figure 2: Word-formation networks generated by the machine learning expansion of the transferred networks, showing the family of lexeme *to reconcile* (encircled) for four of the six language pairs. Top: deu-fra (single lexeme) and fra-deu, middle: ces-fra, bottom: ces-deu.

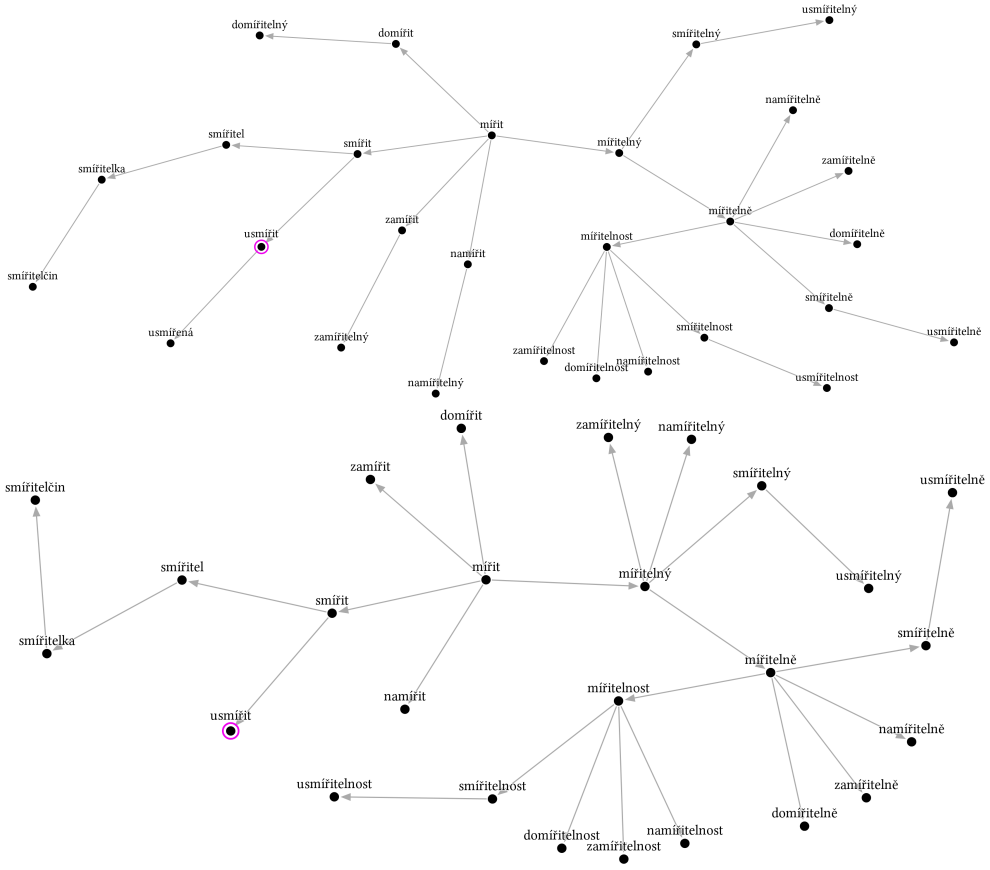


Figure 3: Word-formation networks generated by the machine learning expansion of the transferred networks, showing the family of lexeme *to reconcile* (encircled) for the other two of the six language pairs. Top: deu-ces, bottom: fra-ces.

The machine learning extension method provides a way of generalizing the output of the transfer method, as it learns frequent affixal patterns from the transferred data and applies them to a larger lexicon, omitting infrequent (often spurious) patterns. As seen in the second part of Table 3, this results in increased precision on the networks transferred to French and German, where the gold standard data consists of relatively few selected paradigms and therefore skews towards fewer, more productive patterns. The results on the Czech data, which is more varied, still reach precision comparable to the transferred networks we train on. Recall increases in all cases, even when compared to the "internal" scores, which are more favorable to the transferred networks. Due to this large increase, F1-score also increases. Sample outputs of the machine learning method can be seen in Figures 2 and 3.

The oracle scores for the transfer are in Table 4. The scores are influenced by the ratio of sizes of the word-formation networks used for transfer and evaluation; transferring a large network and evaluating on a smaller one gives an advantage in recall in comparison to the opposite scenario, simply because a larger source network offers more options to select from after transfer. The error causes listed in the table correspond to the sources of error in recall as categorized in Section 5.2.

For all language pairs, most of the errors (50-90%) are attributable to the first cause, where the gold data contains untranslatable lexemes. For the pairs that translate to Czech, this is again explainable by the size and composition of its DeriNet network, which contains many unattested lexemes – finding rare lexemes such as *přeskočitelnost* ("skippability") in the parallel data is unlikely. This is also the reason why the networks obtained through the machine learning expansion have better scores than the oracle of the transfer algorithm. The transfer lexicon is limited to the lexemes found in the parallel data, whose source-side alignments are found in the source word-formation network, and for evaluation purposes, we further limit the lexicon to lexemes from the gold-standard data. The machine-learning pipeline uses the gold-standard lexicon directly, eliminating the "No child trans" class of errors entirely.

Additionally, transfers of networks to German have higher accuracy than transfers to French, even though the recall is comparable. This is because the German network, DErivBase, contains many compounds, which don't have their parents annotated and are listed as unmotivated. These are counted in the accuracy scores (the definition of oracle score above considers missing relations to be always correctly recognized) but do not contribute to recall of relations. The unmotivated words are also the reason behind the fact that the fra-deu pair has higher accuracy than ces-deu, despite having lower recall – fewer relations are translated, resulting in more unmotivated words being correct.

Scores of the cross-lingual embedding model are in Table 5. The model produces results with roughly comparable internal scores (the F1 score on German is 31% for transfer vs. 20% for NNs, while on French it is 27% vs. 39%), but significantly higher gold scores, due to the networks themselves being several times larger. It does not, however, attain scores on par with the machine learning extension method.

The neural extension model exposes a large flaw in the training regime of the neural network. The network is optimized towards minimizing cross entropy between the predicted and gold binary classifications of individual word-formation relations, i.e. it maximizes gold accuracy. As seen in Table 6, the model generally succeeds at that, even though the scores don't necessarily increase on every language (there is a small but significant decrease on Finnish, French and Italian). However, this apparent improvement entirely destroys the usefulness of the model on German and Farsi, because the increase in accuracy is driven by correctly classifying unrelated lexemes at the expense of related ones, causing the recall to go to zero. Even then, the accuracy is still worse than the machine learning extension method. One solution could be a training objective focused on maximizing gold F1 score, or an improved model of word formation which doesn't predict individual relations, but focuses on larger units, e.g. whole instances of a word-formation paradigm or whole word-formation families.

7. Conclusion

In this paper, we presented a two cross-lingual methods for creating word-formation networks – one transfers an existing network using a word-translation lexicon induced from word alignments, the other one uses a neural network with pretrained cross-lingual word embeddings. The transferred small networks are then expanded by either extracting paradigms using statistical machine learning and applying them to a larger set of lexemes, or by bootstrapping the neural network on the small word-formation networks in a cross-lingual fashion. The resulting word-formation networks generally show moderately high precision and good recall.

Acknowledgments

This work was supported by the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19 073/0016935). It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH CZ project (LM2018101).

The OpenSubtitles corpus was kindly provided by `http://www.opensubtitles.org/`.

Bibliography

Baranes, Marion and Benoît Sagot. A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2793–2799, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/379_Paper.pdf`.

Batsuren, Khuyagbaatar, Gabor Bella, and Fausto Giunchiglia. CogNet: A Large-Scale Cognate Database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1302. URL https://www.aclweb.org/anthology/P19-1302.

Batsuren, Khuyagbaatar, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyrool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.89.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X. doi: 10.1162/tacl_a_00051.

Chu, Yoeng-Jin and Tseng-Hong Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965.

Church, Kenneth Ward. Char_align: A Program for Aligning Parallel Texts at the Character Level. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Columbus, Ohio, USA, June 1993. Association for Computational Linguistics. doi: 10.3115/981574.981575. URL https://aclanthology.org/P93-1001.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1073.

Gupta, Prakhar and Martin Jaggi. Obtaining Better Static Word Embeddings Using Contextual Embedding Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5241–5253, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.408. URL https://aclanthology.org/2021.acl-long.408.

Hämmerl, Katharina, Jindřich Libovický, and Alexander Fraser. Combining Static and Contextualised Multilingual Embeddings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2316–2329, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.182. URL https://aclanthology.org/2022.findings-acl.182.

Hathout, Nabil. Acquisition of morphological families and derivational series from a machine readable dictionary. In *Proceedings of the 6th Décembrettes.*, Cascadilla Proceedings Project, pages 166–180, Bordeaux, France, 2008. Cascadilla. URL https://hal.archives-ouvertes.fr/hal-00382808.

Hathout, Nabil and Fiammetta Namer. Démonette, A French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology*, 11:125–162, 2014. doi: 10.33011/lilt.v11i.1369.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4020. URL https://aclanthology.org/P18-4020.

Kyjánek, Lukáš. Morphological Resources of Derivational Word-Formation Relations. Technical Report ÚFAL TR-2018-61, ÚFAL MFF UK, Praha, Czechia, 2018. URL http://ufal.mff.cuni.cz/techrep/tr61.pdf.

Kyjánek, Lukáš, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. Universal Derivations Kickoff: A Collection of Harmonized Derivational Resources for Eleven Languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, pages 101–110, Praha, Czechia, 2019. ÚFAL MFF UK. ISBN 978-80-88132-08-0.

Lango, Mateusz, Zdeněk Žabokrtský, and Magda Ševčíková. Semi-Automatic Construction of Word-Formation Networks. *Language Resources and Evaluation*, 55(1):3–32, 2021. ISSN 1574-020X. doi: 10.1007/s10579-019-09484-2.

McDonald, Ryan, Slav Petrov, and Keith Hall. Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1006.

Musil, Tomáš, Jonáš Vidra, and David Mareček. Derivational Morphological Relations in Word Embeddings. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 173–180, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4818. URL https://aclanthology.org/W19-4818.

Nivre, Joakim et al. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1659–1666. ELRA, 2016.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rosa, Rudolf and Zdeněk Žabokrtský. Unsupervised Lemmatization as Embeddings-Based Word Clustering, 2019.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Simard, Michel, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Montréal, Canada, jun 1992. URL https://aclanthology.org/1992.tmi-1.7.

Straka, Milan and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3009. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Svoboda, Emil and Magda Ševčíková. Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes. *The Prague Bulletin of Mathematical Linguistics*, 118:55–73, 2022. ISSN 0032-6585. doi: 10.14712/00326585.019.

Tiedemann, Jörg. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (*LREC'12*), pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Zeller, Britta, Jan Šnajder, and Sebastian Padó. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1201–1211, Sofia, Bulgaria, 2013. URL http://www.aclweb.org/anthology/P13-1118.

Zhang, Jiajun and Chengqing Zong. Exploiting Source-side Monolingual Data in Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Lan-*

guage Processing, pages 1535–1545, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1160. URL https://aclanthology.org/D16-1160.

Zhang, Yuan, David Gaddy, Regina Barzilay, and Tommi Jaakkola. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1156. URL https://aclanthology.org/N16-1156.

Žabokrtský, Zdeněk, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limburská. Merging Data Resources for Inflectional and Derivational Morphology in Czech. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1307–1314, Paris, France, 2016. European Language Resources Association. ISBN 978-2-9517408-9-1.

Address for correspondence:
Jonáš Vidra
vidra@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics,
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic

The Prague Bulletin of Mathematical Linguistics
NUMBER 120 APRIL 2023

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most interesting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported by some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6–15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive a printed copy of the relevant issue of the PBML together with the original pdf files.

The guidelines for the technical shape of the contributions are found on the web site `https://ufal.mff.cuni.cz/pbml`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.