Delexicalised Multilingual Discourse Segmentation for DISRPT 2021 and Tense, Mood, Voice and Modality Tagging for 11 Languages

Tillmann Dönicke
Göttingen Centre for Digital Humanities
University of Göttingen, Germany
tillmann.doenicke@uni-goettingen.de

Abstract

This paper describes our participating system for the Shared Task on Discourse Segmentation and Connective Identification across Formalisms and Languages. Key features of the presented approach are the formulation as a clause-level classification task, a language-independent feature inventory based on Universal Dependencies grammar, and composite-verb-form analysis. The achieved F1 is 92% for German and English and lower for other languages. The paper also presents a clause-level tagger for grammatical tense, aspect, mood, voice and modality in 11 languages.

1 Introduction

Despite the important role of discourse segmentation for natural language processing (NLP), there is no clear-cut definition of what a discourse segment is. Degand and Simon (2009) determine the boundaries of discourse segments as the intersection of clause boundaries and prosodic boundaries, which means specifically that a discourse segment spans one or several clauses (clauses as minimal discourse segments had been proposed in preceding works, e.g. Mann and Thompson (1988)). We follow this approach and view discourse segmentation as a binary classification problem that predicts for a clause whether it is the start of a new discourse segment. Working with only text makes it impossible to fully implement Degand and Simon (2009)'s approach and include features that capture prosody and prosodic change. Instead, we represent clauses as morphosyntactic feature structures that capture grammatical roles (subject, object etc.), verbal categories and clause connectives, believing that the use of pronouns, the change of tense, aspect and mood, the presence of conjunctions and other linguistic features also signal segment boundaries.

The shared task provides discourse-segmented treebanks for 11 languages. All datasets exist in the Universal Dependencies (UD) format (Nivre et al., 2016). UD grammar (UDG) builds on the idea that all natural languages can be described by a unique inventory of word categories and grammatical rules. Treebanks annotated in UDG thus share the same part-of-speech (POS) tags, morphological features (MFs) and dependency relations (DepRels), which encourages the development of multilingual applications. Things that still significantly differ between languages are the surface forms of words (obviously), the presence/absence of MFs and the order of words and constituents. To alleviate these dissimilarities, we will view sentences as delexicalised, unordered trees and assimilate morphosyntactic features between languages.

Dataset	Sents	Conn.	Delex.	WO
deu.rst.pcc	2,193	no	no	OV
eng.pdtb.pdtb	48,630	yes	yes	VO
eng.rst.gum	8,292	no	yes	–”–
eng.rst.rstdt	8,318	no	yes	–”–
eng.sdrt.stac	11,087	no	no	–”–
eus.rst.ert	2,380	no	no	OV
fas.rst.prstc	2,179	no	no	OV
fra.sdrt.annodis	1,507	no	no	VO
nld.rst.nldt	1,651	no	no	OV
por.rst.cstn	2,221	no	no	VO
rus.rst.rrt	23,044	no	no	VO
spa.rst.rststb	2,089	no	no	VO
spa.rst.sctb	516	no	no	–”–
tur.pdtb.cdtb	31,197	yes	yes	OV
zho.pdtb.cdtb	2,891	yes	yes	VO
zho.rst.sctb	580	no	no	–”–

Table 1: Datasets: total number of sentences, whether discourse connectives are annotated, whether surface forms have been removed, and basic word order.

2 Data and Task

Table 1 gives an overview of the available data. There are 16 datasets for 11 languages, but

Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021), pages 33–45
November 11, 2021. ©2021 Association for Computational Linguistics

the number of sentences for each dataset varies greatly, from 0.5k in `spa.rst.sctb` to 48.6k in `eng.pdtb.pdtb`.

In 13 of the datasets, only discourse segments are annotated: a token which is the begin of a new discourse segment is labelled with `BeginSeg=Yes`. In the remaining three datasets, the full discourse connective is annotated: a token which is the begin of a new discourse segment/connective is labelled with `Seg=B-Conn` and subsequent tokens that are part of the connective are labelled with `Seg=I-Conn`. The task for both annotation schemes is to identify the starts of discourse segments/connectives. In addition, the full connective should be identified for the latter scheme.

Five datasets are only made available in a delexicalised format to participants without a Linguistic Data Consortium membership.

3 Universal Morphosyntactic Features

The following subsections briefly describe how distinct syntactic units are represented in UDG and what features are extracted for the shared task.

3.1 Clauses

Clauses can be extracted from a UDG tree by "cutting" specific clause-marking DepRels. These are: `root`, `csubj`, `ccomp`, `xcomp`, `acl`, `advcl`, `parataxis`, `list`, `vocative` and `discourse`, as well as `conj` if its head is itself governed by a clause-marking DepRel (cf. Dönicke, 2020).[1] Figure 1 shows an example sentence with three clauses, governed by `discourse`, `root` and `xcomp`, respectively. The first two clauses are the starts of a new discourse segment. We handle punctuation at clause boundaries separately. In the example, the comma (,) and the period (.) are stored as preceding and succeeding punctuation of the clause *I 'll try*.

From a clause, we extract the following features: root token's DepRel and POS tag; preceding punctuation; succeeding punctuation.

3.2 NPs

Noun phrases (NPs) realise grammatical roles within a clause. Like clauses, they can be extracted from a UDG tree by cutting specific DepRels. These are: `nsubj`, `obj`, `iobj`, `obl` and `nmod`. In Figure 1, the second clause contains the subject NP *I*, and the third clause contains the object NP *an email*. The **morphological feature structure** (MFS) for each individual word (as given in the data) is shown in (1) and (2), respectively.

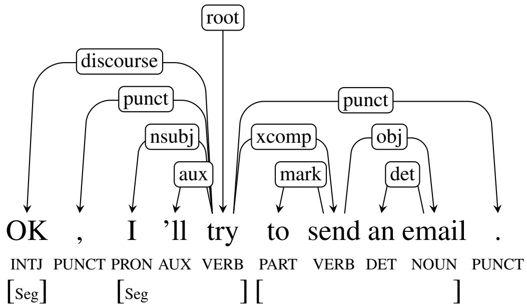
root
discourse
punct
nsubj
aux
xcomp
mark
punct
obj
det
OK , I 'll try to send an email .
INTJ PUNCT PRON AUX VERB PART VERB DET NOUN PUNCT
[Seg] [Seg][ ]

Figure 1: Example sentence from `eng.sdrt.stac`'s development set. Brackets indicate clause spans. All extracted features are shown in Appendix A.

$$\begin{bmatrix} \text{Case} & \text{Nom} \\ \text{Number} & \text{Sing} \\ \text{Person} & 1 \\ \text{PronType} & \text{Prs} \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} \text{Definite} & \text{Ind} \\ \text{PronType} & \text{Art} \end{bmatrix} \begin{bmatrix} \text{Number} & \text{Sing} \end{bmatrix} \tag{2}$$

NP-level features are obtained by unifying the MFSs of the involved words into a single feature structure. As a grammatical rule, Case, Person, Number and Gender have to agree for all words within an NP. Sometimes, this rule is violated (in the data) by compound nouns like *internet problems* where the nouns differ in Number (singular vs. plural). Therefore, we take the agreement features only from the NP's root token; all other features are taken from all words (and are allowed to have multiple values). For a proper handling of analytic languages such as Chinese, which tend to mark features not by morphemes but by particles, we introduce a rule for particles that we apply to an NP's root token w before unifying features:

Particle Rule If w has any particles (i.e. dependents with the POS tag PART), move all particles' features to w and delete the particles.

In Figure 2, for example, the particle *de* [的] has the feature $\begin{bmatrix} \text{Case} & \text{Gen} \end{bmatrix}$, which is moved to the governing noun *lèixíng* [类型].

[1] If the head of a clause-marking DepRel is a modal verb, we do not consider its subtree as a clause because we do not want to separate modal verbs from the verbs they modify; in some treebanks, the modal verb governs the modified verb with an `xcomp` relation (whereas in the English treebanks, the modified verb governs the modal verb with an `aux` relation).

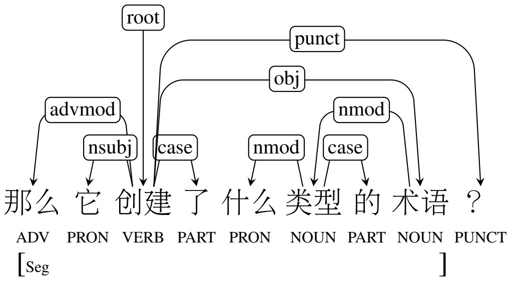
root
punct
obj
advmod
nmod
nsubj
case
nmod
case
那么 它 创建 了 什么 类型 的 术语 ？
ADV PRON VERB PART PRON NOUN PART NOUN PUNCT
[Seg]

Figure 2: Example sentence from `zho.rst.sctb`'s training set.

From each NP in a clause, we extract: root token's DepRel and POS tag; agreement features: Case, Person, Number, Gender; other nominal features: Degree, Definite, Animacy; lexical features: PronType, NumType, Poss, Reflex, Foreign, Abbr, Typo.[2] To make the features NP-specific, we prefix every feature with the root relation, e.g. $\begin{bmatrix}\textsc{nsubj\_Case} & \text{Nom}\end{bmatrix}$, assuming that a clause usually contains only one NP per DepRel.

3.3 Composite Verbs

The composite verb in a clause consists of the main verb and its accompanying full, light, auxiliary and modal verbs as well as verbal particles. (Since we do not distinguish a simple verb form (e.g. *try*) from a compound verb form (e.g. *will try*), we use the term "composite verb" for all cases.) In UDG, we define these as tokens with the POS tag VERB or AUX and subordinate tokens with the POS tag PART and/or the DepRel `compound`. In Figure 1, the second clause contains the composite verb *'ll try*, and the third clause contains the composite verb *to send*. The MFSs of *'ll try* (as given in the data) are shown in (3).

$$\begin{bmatrix}\textsc{VerbForm} & \text{Fin}\end{bmatrix}\begin{bmatrix}\textsc{VerbForm} & \text{Inf}\end{bmatrix} \tag{3}$$

Unfortunately, MFs in the datasets are far from complete; the verbs in (3) are only labelled with VerbForm but not with the other verbal features: Aspect, Mood, Tense, Voice.—An issue that we will take up again in Section 4. For the sake of illustration, we now assume that the MFs are complete, as shown in (4). Note that English finite verbs do not mark Aspect and Voice at the morphological level and English infinitives do not have any inflectional features (both properties differ in other languages).

$$\begin{bmatrix}\textsc{Tense} & \text{Pres}\\ \textsc{Mood} & \text{Ind}\\ \textsc{VerbForm} & \text{Fin}\end{bmatrix}\begin{bmatrix}\textsc{VerbForm} & \text{Inf}\end{bmatrix} \tag{4}$$

Combining the MFSs of the individual words into a single feature structure is not as easily possible as for NPs since there are no linguistic unification/agreement rules amongst the words in a composite verb as they exist for NPs. A simple method for feature extraction would still be to use MFs and prefix them by the POS tag of the corresponding word (and allowing multiple values if there are more than one words with the same POS tag). The **morphosyntactic feature structure** (MSFS) resulting from (4) is shown in (5).

$$\begin{bmatrix}\text{AUX\_}\textsc{Mood} & \text{Ind}\\ \text{AUX\_}\textsc{Tense} & \text{Pres}\\ \text{AUX\_}\textsc{VerbForm} & \text{Fin}\\ \text{VERB\_}\textsc{VerbForm} & \text{Inf}\end{bmatrix} \tag{5}$$

However, grammaticalised composite verb constructions are quite different for the languages of the world (and also for those in the shared task). Another way to represent (4) is as the **grammatical feature structure** (GFS) in (6).

$$\begin{bmatrix}\textsc{Aspect} & \text{Imp}\\ \textsc{Mood} & \text{Ind}\\ \textsc{Tense} & \text{Fut}\\ \textsc{VerbForm} & \text{Fin}\\ \textsc{Voice} & \text{Act}\end{bmatrix} \tag{6}$$

Arriving at grammatical features (GFs) is a complex task on its own, which is why we describe the procedure separately in Section 4. Note that the structure in (6) includes $\begin{bmatrix}\textsc{Tense} & \text{Fut}\end{bmatrix}$, since *will try* is grammatically future tense, whereas (5) only includes $\begin{bmatrix}\text{AUX\_}\textsc{Tense} & \text{Pres}\end{bmatrix}$ because of the morphological present tense of *will*. GFs assimilate universal clause representations in such that they encode *which* features are expressed by a composite verb and not *how* the verbs are composed. For example, most languages have grammatical future tense, but in some languages (e.g. English) future tense is only marked grammatically whereas in others (e.g. Basque) it is also marked morphologically. We thus assume that GFSs show a greater similarity between languages than MSFSs. Note, however, that GFSs still exhibit differences between languages, because not all languages have parallel grammaticalised constructions.[3]

[2] Explanations and possible values for all of these features can be found at https://universaldependencies.org/u/feat/.

[3] Just to give an example: English has progressive aspect which German has not. The same holds for NPs: German has dative case which English has not.

3.4 Free Discourse Elements

Some words are neither part of an NP nor of the composite verb. If these words are clause-level, i.e. directly governed by the clause's root token, we call them "free discourse elements". These elements comprise e.g. adverbs, complementisers and conjunctions, and are thus very interesting for the task of discourse segmentation. Therefore, we extract DepRel and POS tag from every free discourse element. As for NP-level features, we prefix every feature with the element's DepRel, e.g. $\begin{bmatrix}\textsc{mark\_pos} & \text{SCONJ}\end{bmatrix}$.

3.5 Feature Vectors

When vectorising a document $D = [c_1, \dots, c_n]$, we get clause vectors $\vec{c}_1, \dots, \vec{c}_n$, which we then concatenate to context-sensitive vectors $X_D = [\vec{x}_1, \dots, \vec{x}_n]$ using a window of 3 clauses: $\vec{x}_i = \vec{c}_{i-1} \circ \vec{c}_i \circ \vec{c}_{i+1}$. For the context clauses c_{i-1} and c_{i+1}, we add additional features that indicate whether the clause is in the same sentence as c_i and whether the clause is directly subordinate or directly superordinate to c_i. The classes corresponding to X_D are $Y_D = [y_1, \dots, y_n]$ with $y_i \in \{\text{TRUE}, \text{FALSE}\}$ (see Section 5.1).

For the documents that include discourse connectives, we create additional vectors $X_D^{Conn} = [\vec{d}_1, \dots, \vec{d}_m]$ for the connectives $d_1, \dots, d_m$. Let c_j be the clause that starts with d_j. To construct $\vec{d}_j$, we extract from the first 5 tokens of c_j: POS tag; DepRel; index (starting at 1) of head if the head is among the first five tokens, 0 otherwise. (All features are index-specific, e.g. $\begin{bmatrix}\text{1\_POS} & \text{INTJ}\end{bmatrix}$.) Since not every clause contains 5 or more tokens, we further add a feature with value $\min\{|c_j|, 5\}$.[4] We will use these features to predict the length of the discourse connectives $Y_D^{Conn} = [|d_1|, \dots, |d_m|]$ (see Section 5.2).

4 Grammatical TMVM Tagging

Dönicke (2020) presents an algorithm for tagging the GFs Tense, Aspect, Mood, Voice and Modality (TMVM) of a clause in German. The algorithm identifies the words that contribute to a composite verb and uses a function R that maps a bag of MFSs to a GFS, like

$$R\left(\left\{\begin{bmatrix}\textsc{lemma} & \textit{will}\\ \textsc{Tense} & \text{Pres}\\ \textsc{Mood} & \text{Ind}\\ \textsc{VerbForm} & \text{Fin}\end{bmatrix}, \begin{bmatrix}\textsc{VerbForm} & \text{Inf}\end{bmatrix}\right\}\right) = \begin{bmatrix}\textsc{Aspect} & \text{Imp}\\ \textsc{Mood} & \text{Ind}\\ \textsc{Tense} & \text{Fut}\\ \textsc{VerbForm} & \text{Fin}\\ \textsc{Voice} & \text{Act}\end{bmatrix},$$

where R relies on a comprehensive table of all composite verb constructions (i.e. the complete conjugation table of the language). Note that only the lemmas of auxiliary verbs are relevant for the algorithm since R does not depend on the main verb. In addition to a list of auxiliary verbs, a list of modal verbs is required.[5]

Algorithm 1 shows an updated version of the original algorithm that has been modified to work with a broader range of languages, specifically the languages in the shared task. In the following, the algorithm is briefly described, with a focus on the adaptions made for multiple languages (numbers in parentheses refer to lines in the pseudocode); for further explanations see Dönicke (2020).

Given a composite verb $V = [v_1, \dots, v_{|V|}]$ in a language ℓ, first of all the particle rule from Section 3.2 is applied to all words (ll. 1–2). Considering the Chinese example in Figure 2 again, this moves the feature $\begin{bmatrix}\textsc{Aspect} & \text{Perf}\end{bmatrix}$ from the particle *le* [了] to its governing verb *chuàngjiàn* [创建] and removes the particle from V. After this step, V contains only verbs.[6]

The algorithm is designed for an OV language, i.e. a language in which the basic order of object (O) and verb (V) is O-V. More importantly for the algorithm, the basic order of auxiliary (Aux) and verb in OV languages is V-Aux, whereas it is Aux-V in VO languages (Dryer, 1992). Thus, if the input language ℓ is a VO language (see Table 1), V has to be reversed before going on (ll. 3–4).

To counteract finite-verb movement in some languages (e.g. German and Dutch), finite verbs and non-finite verbs are selected separately (ll. 5–6) and then the finite verb is inserted at the syntactically highest position (ll. 7–9). After this step, all verbs in V should be ordered from syntactically lowest

[4] 0.3% of the discourse segments start with a connective that is longer than 5 tokens. These connectives are ignored.

[5] The algorithm differentiates between full verbs, auxiliary verbs and modal verbs. However, UDG only distinguishes full verbs (VERB) and auxiliary verbs (AUX), and modal verbs are tagged as either, depending on the language (see https://universaldependencies.org/u/pos/AUX_.html). Therefore, a list of modal verbs is required to identify them. As a reviewer pointed out, one could also use the language-specific POS tags to identify modal verbs. In our implementation, we also need the lemmas for modal unification across languages (see text).

[6] As in linguistic works (e.g. Antonenko, 2008; Dobrushina, 2012), we also consider the Russian бы as particle although it is tagged with AUX in the Russian dataset.

Algorithm 1: Compute features of composite verb V in language ℓ

```
1  for $i = 1$ to $|V|$ do
2      particle_rule($v_i$)
3  if $\ell$ is VO language then
4      $V \leftarrow [v_{|V|}, \ldots, v_1]$
5  $V_{fin} \leftarrow$ [finite verbs in $V$]
6  $V \leftarrow$ [non-finite verbs in $V$]
7  if $|V_{fin}| > 0$ then
8      $v_{fin} \leftarrow$ right-most finite verb in $V_{fin}$
9      $V \leftarrow [v_1, \ldots, v_{|V|}, v_{fin}]$
10 if $|V| = 0$ then
11     return [ ]
12 else if main verb in $V$ then
13     $v_{main} \leftarrow$ right-most main verb in $V$
14 else
15     $v_{main} \leftarrow$ left-most verb in $V$
16 $V \leftarrow [v_{main}, \ldots, v_{fin}]$
17 $M \leftarrow$ [features*$(v_i, \ell)$ for $i = 1$ to $|V|$]
18 for $i = |V|$ to $1$ do
19     if $v_i$ is modal verb then
20         $m_{i-1} \leftarrow m_i$
21 while $|V| > 0$ do
22     Set $v_1$ to be the main verb
23     $F \leftarrow \bigtimes_{\substack{1 \leq i \leq |V| \\ v_i \text{ is not modal verb}}} m_{|M|-|V|+i}$
24     if $\sum_{i=1}^{|F|} \sum_{j=1}^{|f_i|} |f_{ij}| = 0$ then
25         return [ ]
26     $A \leftarrow \{\}$
27     for $i = 1$ to $|F|$ do
28         $A \overset{\cup}{\leftarrow} R^*(f_i, \ell)$
29     if $|A| = 0 \land |V| = 1$ then
30         $A \leftarrow m_{|M|}$
31     if $|A| > 0$ then
32         $A \leftarrow$ filter($A$)
33         $a \leftarrow$ combine($A$)
34         $a \leftarrow$ unify_verb_form($a$)
35         $V_{modal} \leftarrow$ [modal verbs in $V$]
36         $V_{modal} \leftarrow$ unify_modals($V_{modal}, \ell$)
37         $a \overset{\sqcup}{\leftarrow} \begin{bmatrix}\textsc{Modality} & V_{modal}\end{bmatrix}$
38         return $a$
39     $V \leftarrow [v_2, \ldots, v_{|V|}]$
40 return [ ]
```

$\overset{\cup}{\leftarrow}$ and $\overset{\sqcup}{\leftarrow}$ are augmented assignment operators for union and unification, respectively.

to highest position.

If V is not empty (ll. 10–11), the main verb is determined (ll. 12–15) and all syntactically lower verbs are removed from V (l. 16), because they are not relevant for TMVM tagging.

The MFs of each $v_i \in V$ are stored in $m_i \in M$ (l. 17), where m_i is a set of MFSs since it is (theoretically) possible that a verb is morphologically ambiguous. However, since the MFs for each verb are given in the data, $|m_i| = 1$ per default.[7,8]

As in the original algorithm, MFs of modal verbs overwrite those of syntactically lower verbs (ll. 18–20). All possible combinations of the involved verbs' MFSs, excluding modal verbs, are then stored in $F = \{f_1, \ldots, f_{|F|}\}$ (l. 23). In a simple case with no modal verbs and $|m_i| = 1$ for all $m_i \in M$, $|F| = 1$ and f_1 contains the MFS of every verb $v_1, \ldots, v_{|V|}$, i.e. $f_1 = \{m_{11}, \ldots, m_{|V|_1}\}$.[9]

Every combination $f_i \in F$ is then analysed with the language-specific look-up table and the analyses (i.e. GFSs) are stored in A (ll. 26–28).[10] As mentioned in Section 3.3, a lot of MFs are missing in the data. R^* treats missing features as features with wildcard values and returns all matching analyses, which means that the number of returned analyses for f_i increases with the number of missing features in each $f_{ij} \in f_i$ and would become maximal if every f_{ij} is empty. As a basic restriction, we require that at least one f_{ij} is not empty and return an empty feature structure otherwise (ll. 24–25).

In contrast to too many analyses, it is also possible that no analysis is found. In this case, the syntactically lowest verb is removed (l. 39) and the look-up is repeated (l. 21). If only one verb is left and still no analysis is found, A is set to the verb's MFSs (ll. 29–30).

The analyses in A are then filtered (l. 32), e.g. by giving higher preference to analyses with

[7] As in Dönicke (2020), we add the participle analysis to potential substitute infinitives in German.

[8] We perform a small number of modifications to the MFs for cases where we think that the data is not labelled ideally. For example, some languages use $\begin{bmatrix}\textsc{VerbForm} & \text{Ger}\end{bmatrix}$ and some use $\begin{bmatrix}\textsc{Tense} & \text{Pres} \\ \textsc{VerbForm} & \text{Part}\end{bmatrix}$ for very similar forms of the verb (gerunds and present participles). For this reason, the UD guidelines discourage the use of $\begin{bmatrix}\textsc{VerbForm} & \text{Ger}\end{bmatrix}$ (see https://universaldependencies.org/u/feat/VerbForm.html#Ger) and we convert it to the latter feature combination.

[9] Actually, the Cartesian product yields an ordered combination $[m_{11}, \ldots, m_{|V|_1}]$ but we treat it as unordered combination to be less prone to potential local verb movements.

[10] The look-up tables have been created manually. For some languages, this required extensive study of composite verb constructions, and we want to acknowledge a few works that were very helpful in this process: Berro et al. (2019) for Basque, Izadi and Rahimi (2015) for Persian, Babby and Brecht (1975) for Russian, Jendraschek (2011) for Turkish, and Li and Thompson (1989) for Chinese.

$\begin{bmatrix}\textsc{Voice} & \text{Act}\end{bmatrix}$ and/or $\begin{bmatrix}\textsc{Mood} & \text{Ind}\end{bmatrix}$. The remaining analyses are unified into a single GFS a, ignoring features with conflicting values (l. 33).

Since not all languages have the same types of non-finite verb forms, we normalise them as follows (l. 34):

$$\begin{bmatrix}\textsc{VerbForm} & \text{Inf}\end{bmatrix} \sqsubseteq a : a \leftarrow \begin{bmatrix}\textsc{VerbForm} & \text{Inf}\\ \textsc{VerbForm*} & \text{Verb}\end{bmatrix}$$

$$\begin{bmatrix}\textsc{VerbForm} & \text{Vnoun}\end{bmatrix} \sqsubseteq a : a \leftarrow \begin{bmatrix}\textsc{VerbForm} & \text{Inf}\\ \textsc{VerbForm*} & \text{Noun}\end{bmatrix}$$

$$\begin{bmatrix}\textsc{VerbForm} & \text{Part}\end{bmatrix} \sqsubseteq a : a \leftarrow \begin{bmatrix}\textsc{VerbForm} & \text{Part}\\ \textsc{VerbForm*} & \text{Adj}\end{bmatrix}$$

$$\begin{bmatrix}\textsc{VerbForm} & \text{Conv}\end{bmatrix} \sqsubseteq a : a \leftarrow \begin{bmatrix}\textsc{VerbForm} & \text{Part}\\ \textsc{VerbForm*} & \text{Adv}\end{bmatrix}$$

In a last step, we add the modal verbs to a (ll. 35–37). In Dönicke (2020), the lemmas of the verbs are used but in our multilingual implementation, we map the lemmas to three categories of modal verbs (cf. Biber et al., 2002, p. 176): permission/possibility/ability (POS), obligation/necessity (OBL), and volition/prediction (VOL).

5 Classification

We expect a high interdependence between the extracted features, which is why we use decision trees for the classification of clauses. A decision tree is a statistical classification method that can both learn such complex dependencies and also visualise them in an understandable manner.

Experiments with other classifiers, including complement Naive Bayes, random forest and multilayer perceptron, could not improve the performance over that of a simple decision tree. This suggests that the decision tree makes the best out of the available features.

5.1 Discourse Segments

Given a training set $X_{D_{train}}$, we train a decision tree classifier with Gini impurity as split criterion. Since the performance of a decision tree strongly depends on its depth and leaf size, grid search is performed to select the optimal values for the maximum tree depth in $\{5, 10, 15, 20, 25, \infty\}$ and the minimum leaf size in $\{1, 2, 5, 10, 15, 20\}$. For the grid search, the development set $X_{D_{dev}}$ corresponding to $X_{D_{train}}$ is used for validation.

5.2 Discourse Connectives

The classifier that predicts the length of a connective is also a decision tree with Gini impurity as split criterion. We let this tree fully expand on the training set $X^{Conn}_{D_{train}}$ (maximum tree depth = ∞; minimum leaf size = 1) since we assume that discourse connectives are like a closed class and generalising to unseen feature combinations is rarely needed.

6 Experiments

Parsed vs. Plain As suggested in the shared task, we evaluate our systems in two main conditions: using the parsed/treebanked datasets (`.conllu` files) and using the plain/tokenised datasets (`.tok` files). We approach the second condition by preprocessing the plain datasets with spaCy (https://spacy.io/) and training new classifiers on the processed training sets. SpaCy provides pretrained UDG models for all shared task's languages except German, Persian, Basque and Turkish. For these languages, we trained new models on the UD treebanks HDT (German), PerDT (Persian), BDT (Basque) and Kenet (Turkish) (Zeman et al., 2021).

Morphosyntactic vs. Grammatical In all experiments, we represent composite verbs either as morphosyntactic (M) or as grammatical (G) feature structures (as described in Section 3.3).

Monolingual vs. Multilingual Each system is evaluated on all 16 test sets. In the monolingual condition, we train a system on one dataset only. We further train a system on all training sets combined (ALL) as well as 16 systems on all but one training sets (CV). In the CV condition, we evaluate the system on the test set corresponding to the excluded training set. Thus, the CV condition corresponds to a scenario without training data for the test language.

7 Results and Discussion

Tables 2 and 3 show the results in the parsed and the plain condition. Numbers are the F1 scores for discourse segmentation or connective identification, depending on the dataset. For the monolingual experiments, the highest value in each column is boldfaced. For the multilingual experiments, the higher value on each test set is underlined. In the monolingual experiments, the F1 scores for parsed data are on average 3.5% higher than those for plain data. The best result on a test set is usually achieved by the system trained on the corresponding training set.

The systems presented in this paper do not perform better than the best systems from DISRPT

	Test set																															
	deu. rst. pcc		eng. pdtb. pdtb		eng. rst. gum		eng. rst. rstdt		eng. sdrt. stac		eus. rst. ert		fas. rst. prstc		fra. sdrt. annodis		nld. rst. nldt		por. rst. cstn		rus. rst. rrt		spa. rst. rststb		spa. rst. sctb		tur. pdtb. tdb		zho. pdtb. cdtb		zho. rst. sctb	
Training set	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G
deu.rst.pcc	**89**	**92**	26	25	73	76	66	62	91	91	71	71	66	63	57	58	87	89	67	72	64	65	80	78	65	63	7	7	18	18	63	65
eng.pdtb.pdtb	29	27	**74**	**74**	29	29	25	21	8	8	25	25	20	21	13	14	22	21	22	21	24	24	25	26	16	15	13	16	7	10	7	8
eng.rst.gum	88	89	26	28	**86**	**86**	**80**	**80**	89	89	69	70	74	73	60	61	85	85	74	74	67	67	74	74	57	57	6	6	19	**21**	52	55
eng.rst.rstdt	86	87	26	26	83	83	75	75	88	88	69	70	75	69	61	61	83	83	73	72	63	65	70	72	55	55	6	6	19	19	43	48
eng.sdrt.stac	85	85	21	21	64	64	51	51	**92**	**92**	70	70	61	61	50	50	82	82	64	64	59	59	77	77	68	68	8	8	16	16	73	**73**
eus.rst.ert	86	86	28	27	65	66	55	53	91	90	**73**	**75**	64	62	52	52	85	86	67	65	65	64	80	80	67	67	12	9	17	17	**74**	69
fas.rst.prstc	87	88	34	32	72	73	64	68	91	90	**73**	71	**80**	**81**	54	56	83	83	68	70	67	67	74	77	61	62	7	8	**20**	19	68	65
fra.sdrt.annodis	87	87	27	32	76	78	55	71	72	88	69	70	70	71	**62**	**62**	85	86	73	73	66	66	75	76	58	59	7	7	16	18	72	53
nld.rst.nldt	87	87	32	35	74	73	58	57	91	90	72	74	67	67	55	57	**90**	**90**	73	69	68	67	83	83	66	64	11	12	17	17	64	72
por.rst.cstn	88	86	30	33	76	73	59	23	90	59	69	69	63	65	60	57	85	84	**80**	**80**	67	66	78	76	62	60	10	7	18	3	59	56
rus.rst.rrt	**89**	89	26	26	69	70	55	54	85	86	71	70	62	66	55	55	86	87	71	71	**73**	**73**	78	79	63	64	9	9	17	16	68	68
spa.rst.rststb	87	87	30	32	69	70	54	54	83	83	69	69	56	58	54	52	79	83	70	69	66	66	**86**	**85**	**69**	**69**	8	7	16	16	62	62
spa.rst.sctb	78	85	20	27	60	67	51	51	81	89	65	68	56	19	48	48	76	80	63	65	59	60	79	78	65	66	8	7	16	16	73	39
tur.pdtb.tdb	5	7	30	33	11	14	8	8	1	1	10	10	9	14	8	8	9	12	14	15	12	14	11	12	10	10	**37**	**37**	1	1	2	2
zho.pdtb.cdtb	0	0	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0
zho.rst.sctb	69	69	17	17	53	53	47	47	65	65	64	64	53	53	46	46	72	72	54	54	56	56	73	73	62	62	9	9	15	15	72	72
ALL	89	89	72	71	77	77	26	20	91	91	73	73	76	76	59	58	89	88	76	76	72	73	84	84	65	67	34	33	8	10	70	68
CV	90	90	26	26	71	70	24	19	81	83	72	71	67	63	58	57	88	87	74	74	65	66	81	80	66	66	13	13	5	13	65	62

Table 2: Results on the parsed data in %.

	Test set																															
	deu. rst. pcc		eng. pdtb. pdtb		eng. rst. gum		eng. rst. rstdt		eng. sdrt. stac		eus. rst. ert		fas. rst. prstc		fra. sdrt. annodis		nld. rst. nldt		por. rst. cstn		rus. rst. rrt		spa. rst. rststb		spa. rst. sctb		tur. pdtb. tdb		zho. pdtb. cdtb		zho. rst. sctb	
Training set	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G	M	G
deu.rst.pcc	**89**	**89**	–	–	–	–	–	–	53	53	69	69	71	71	57	56	82	81	67	67	62	62	71	73	63	63	–	–	–	–	65	64
eng.pdtb.pdtb	–	–	*48*	*48*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
eng.rst.gum	–	–	–	–	*75*	*75*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
eng.rst.rstdt	–	–	–	–	–	–	*80*	*80*	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
eng.sdrt.stac	78	83	–	–	–	–	–	–	**67**	**69**	60	62	62	71	55	62	79	81	68	72	58	58	72	70	59	57	–	–	–	–	55	48
eus.rst.ert	83	84	–	–	–	–	–	–	51	52	**72**	**73**	55	56	53	53	82	81	63	67	60	59	74	76	69	67	–	–	–	–	69	69
fas.rst.prstc	88	86	–	–	–	–	–	–	57	57	70	69	**77**	**74**	57	56	83	83	67	66	64	63	74	74	63	65	–	–	–	–	67	67
fra.sdrt.annodis	85	85	–	–	–	–	–	–	59	55	65	65	65	64	**66**	**65**	83	83	73	76	64	65	71	72	57	61	–	–	–	–	54	64
nld.rst.nldt	86	85	–	–	–	–	–	–	54	53	71	71	66	64	56	57	**87**	**87**	67	68	64	64	77	77	62	**68**	–	–	–	–	67	**70**
por.rst.cstn	84	85	–	–	–	–	–	–	54	54	68	67	64	61	60	62	84	85	**77**	**77**	65	65	74	72	62	59	–	–	–	–	69	69
rus.rst.rrt	87	87	–	–	–	–	–	–	48	51	70	68	61	65	56	58	80	83	68	72	**71**	**71**	73	73	64	64	–	–	–	–	68	68
spa.rst.rststb	84	82	–	–	–	–	–	–	43	46	68	69	57	64	55	54	79	80	69	68	63	61	**80**	**81**	68	**68**	–	–	–	–	**70**	**70**
spa.rst.sctb	85	82	–	–	–	–	–	–	51	46	65	65	59	59	53	53	81	77	64	63	60	55	77	78	**69**	**68**	–	–	–	–	**70**	**70**
tur.pdtb.tdb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	*41*	*41*	–	–	–	–
zho.pdtb.cdtb	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	*9*	*9*	–	–
zho.rst.sctb	80	80	–	–	–	–	–	–	50	50	63	63	56	56	45	45	77	77	59	59	55	55	71	71	68	**68**	–	–	–	–	69	69
ALL	86	88	–	–	–	–	–	–	65	65	72	72	73	74	61	62	84	85	75	75	71	71	80	79	67	65	–	–	–	–	67	69
CV	85	87	–	–	–	–	–	–	49	51	70	70	63	63	61	59	84	83	72	73	65	65	76	78	68	65	–	–	–	–	65	65

Table 3: Results on the plain data in %. Delexicalised datasets are excluded because they cannot be preprocessed; italicised results have been obtained by the shared task organisers.

2019 (Zeldes et al., 2019). A fundamental difference between previous systems and the current system is the classification approach: Whereas previous works performed a token-level classification, the current work tries a clause-level classification. The latter approach relies on the assumption that starts of discourse segments are almost always starts of clauses; and it was our mistake and maybe also bit of an unfortunate coincidence that we checked this hypothesis only for deu.rst.pcc and eng.sdrt.stac (the datasets which we mostly used for development), where indeed 95% and 97%, respectively, of the segment starts coincide with clause starts. As we can see in Table 4, the percentage is much lower in other datasets. Since we train our decision tree to

Dataset	%	R	P	F1	F1$^*_1$	F1$^*_2$
discourse segmentation						
deu.rst.pcc	95	89	95	92	94	–
eng.rst.gum	85	80	94	86	94	–
eng.rst.rstdt	84	66	87	75	80	–
eng.sdrt.stac	97	86	99	92	94	–
eus.rst.ert	85	66	87	75	84	–
fas.rst.prstc	80	74	88	81	91	–
fra.sdrt.annodis	65	50	82	62	83	–
nld.rst.nldt	93	86	95	90	95	–
por.rst.cstn	74	68	97	80	95	–
rus.rst.rrt	82	64	86	73	85	–
spa.rst.rststb	91	82	89	85	90	–
spa.rst.sctb	88	57	79	66	79	–
zho.rst.sctb	88	67	79	72	81	–
mean	85	72	89	79	88	–
connective identification						
eng.pdtb.pdtb	82	66	86	74	84	97
tur.pdtb.tdb	44	24	80	37	63	96
zho.pdtb.cdtb	40	1	30	2	6	89
mean	55	30	65	38	51	94

Table 4: Percentage of segment starts that are also clause starts, and achieved recall, precision and F1 (see Table 2) for each dataset. F1$^*_1$ and F1$^*_2$ are the F1 scores of the individual classifiers that predict clause-initial segment starts and connective lengths, respectively.

distinguish clauses where the first token is the start of a discourse segment from all other clauses (including clauses that contain starts of discourse segments at non-initial positions), the percentage sets an upper bound for the classification recall. The languages with the highest achieved precision are English (86%–99%), Portuguese (97%), German (95%) and Dutch (95%); the languages with the highest F1 are German (92%), English (74%–92%) and Dutch (90%). If only clause-initial segment starts are taken into account (F1$^*_1$ in Table 4), the F1 of the decision tree significantly increases for almost all datasets (+9% on average). The performance for determining the length of discourse connectives ranges between 89% and 97% (F2$^*_1$ in Table 4).

These results suggest that the clause-level approach could achieve reasonable results if segment starts would always coincide with clause starts. This precondition is, however, hard to fulfil, since there are not only different frameworks for discourse segmentation but also different notions of what a clause is. In this paper, we define clauses in terms of UDG. In practice, UD annotations are carried out by many different research groups or converted from non-UD treebanks and thus prone to inconsistencies that may also affect the annotations of clause-marking relations (e.g. de Marneffe et al., 2017). Furthermore, a lot of the datasets in the shared task incorporate automatically created dependency trees (created by models trained on UD treebanks), which may lead to follow-up errors in the clause-splitting step. Dönicke (2020) reports an F1 of 81% for predicting clauses in a German text after preprocessing it with a spaCy model trained on the German UD treebanks. Even though this number only gives a rough estimate on how well our system identifies clauses, there is clearly room for improvement. One could also try to resolve the mismatch between segment starts and clause starts in a postprocessing step, e.g. by a second classifier that identifies the position of a segment start in a clause (similar to our connective-length classifier).

All of the systems from DISRPT 2019 use lexical features, where the best systems (Muller et al., 2019; Iruskieta et al., 2019) are recurrent neural networks. The system that is most similar to the current work is the (best) system from Bourgonje and Schäfer (2019), who use a random forest classifier and extract features at the token-level, e.g. surface form, POS tag, position in the sentence, succeeding punctuation mark. Like we do with clauses, they extract features from the current, the preceding and the succeeding token. For German and Basque, our clause-level, delexicalised and unordered-tree approach yields higher F1s than Bourgonje and Schäfer (2019)'s random forest; these are, however, the only languages on which our system performs better. The motivation for not using lexical features was to create language-independent, universal representations for multilingual learning. However, lexical features potentially improve the performance in a monolingual setting.[11]

[11] A multilingual alternative to lexical features are semantic features, which we also experimented with in the development phase. We extracted semantic features for English verbs and their synonyms in the other languages from ConceptNet (Speer et al., 2017), and added the features of a clause's main verb to its grammatical feature structure. (The most common semantic features are: change, contact, communication, motion, social, stative, possession, cognition, body, creation, perception, emotion.) Using these features could not improve our results. A possible reason for this is that we assigned semantic features without disambiguating verb senses and therefore a lot of verbs received a broad range of features. However, we are not aware of an existing multilingual resource for word

Training on all languages does never improve over the performance of the best monolingual system. The results with a multilingual training set, however, might be distorted because `eng.pdtb.pdtb`, `tur.pdtb.tdb` and `rus.rst.rrt` constitute far larger parts in the multilingual training set than the other datasets. This is also visible in the CV experiments: when one of the large datasets is excluded, the performance drops more than when a smaller dataset is excluded. For example, the performance drops from 71% to 26% when `eng.pdtb.pdtb` is excluded, whereas it drops from 67% to 66% when `spa.rst.sctb` (the smallest dataset) is excluded. Although our system does not profit from multilingual training in the context of the shared task, it might be useful in scenarios where no training data is available for a language. For example, training and testing on `spa.rst.rststb` achieves an F1 of 85%, and training on other RST treebanks leads to 72%–83% on `spa.rst.rststb` (see Table 2). Note that training on the small but same-language treebank `spa.rst.sctb` yields 78%, whereas training on `nld.rst.nldt`, which is three times as large, yields 83%. Joining only some and not all datasets might improve the performance for individual languages as well. Future work on multilingual training could also experiment with balanced datasets.

The use of GFSs instead of MSFSs did not have a great impact on the classification performance (<1% difference on average). Inspection of the learned decision trees showed that the top-level features concern punctuation, clause types, free discourse elements and partially NPs, but features concerning the verb are less common. (As an extreme example, the decision tree for `eng.sdrt.stac`, see Appendix B, does not include any verbal feature.) Unexpectedly, tense, mood, voice etc. seem to be irrelevant for discourse segmentation, and so it does not matter how they are represented.

8 Conclusion

In this paper, we approached discourse segmentation as a clause-level classification task and represented clauses as delexicalised UD-based feature structures. While the approach works sufficiently on some datasets (e.g. German), the performance is generally lower than that of other approaches (cf. Zeldes et al., 2019). A major reason for this is that, contrary to our expectation, boundaries of discourse segments do not typically fall onto clause boundaries in most datasets.

In the context of the shared task, we extended Dönicke (2020)'s algorithm for the grammatical analysis of composite verb forms and created the language-specific resources to run it for all 11 languages. Thus, we also contribute to the task of compound-verb analysis, which is (in contrast to morphological analysis) underrepresented in NLP.[12] However, annotating data with grammatical features and testing the algorithm goes beyond the scope of participating in the shared task and is left to future work.

Our system is available at https://gitlab.gwdg.de/tillmann.doenicke/disrpt2021-tmvm.

References

Andrei Antonenko. 2008. The nature of Russian subjunctive clauses.

Leonard H. Babby and Richard D. Brecht. 1975. The syntax of voice in Russian. *Language*, 51(2):342–367.

Ane Berro, Fernández Beatriz, and Jon Ortiz de Urbina, editors. 2019. *Basque and Romance: Aligning Grammars*. Grammars and Sketches of the World's Languages. Brill.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 2002. *Longman grammar of spoken and written English*. Second impression 2003.

Peter Bourgonje and Robin Schäfer. 2019. Multilingual and cross-genre discourse unit segmentation. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 105–114, Minneapolis, MN. Association for Computational Linguistics.

Narayan Choudhary, Pramod Pandey, and Girish Nath Jha. 2014. A rule based method for the identification of TAM features in a PoS tagged corpus. In *Human Language Technology Challenges for Computer Science and Linguistics*, pages 178–188, Cham. Springer International Publishing.

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the Universal Dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115, Pisa,Italy. Linköping University Electronic Press.

sense disambiguation and semantic feature assignment.

[12] We are aware of works on Czech (Žáčková et al., 2000), Hindi (Choudhary et al., 2014), Italian (Faro and Pavone, 2015), German, French and English (Ramm et al., 2017; Myers and Palmer, 2019).

Liesbeth Degand and Anne Catherine Simon. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours* [En ligne], 4.

Nina Dobrushina. 2012. Subjunctive complement clauses in Russian. *Russian linguistics*, 36(2):121–156.

Tillmann Dönicke. 2020. Clause-level tense, mood, voice and modality tagging for German. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 1–17, Düsseldorf, Germany. Association for Computational Linguistics.

Matthew S. Dryer. 1992. The Greenbergian word order correlations. *Language*, 68:81–138.

Simone Faro and Arianna Pavone. 2015. Refined tagging of complex verbal phrases for the Italian language. In *Proceedings of the Prague Stringology Conference 2015*, pages 132–145, Czech Technical University in Prague, Czech Republic.

Mikel Iruskieta, Kepa Bengoetxea, Aitziber Atutxa Salazar, and Arantza Diaz de Ilarraza. 2019. Multilingual segmentation based on neural networks and pre-trained word embeddings. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 125–132, Minneapolis, MN. Association for Computational Linguistics.

Mehri Izadi and Maryam Rahimi. 2015. Word order of Persian and English: A processing-based analysis. *Education Journal*, 4(1):37–43.

Gerd Jendraschek. 2011. A fresh look at the tense-aspect system of Turkish. *Language Research*, 47(2):245–270.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Skatje Myers and Martha Palmer. 2019. ClearTAC: Verb tense, aspect, and form classification using neural nets. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 136–140, Florence, Italy. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Anita Ramm, Sharid Loáiciga, Annemarie Friedrich, and Alexander Fraser. 2017. Annotating tense, mood and voice for English, French and German. In *Proceedings of ACL 2017, System Demonstrations*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Eva Žáčková, Luboš Popelínský, and Miloslav Nepil. 2000. Automatic tagging of compound verb groups in Czech corpora. In *Text, Speech and Dialogue*, pages 115–120, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, **H**órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu,

Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon. Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, Kyung-Tae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Rachele Sprugnoli, Stein**h**ór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa,

Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Clause Representations

The three clauses from Figure 1 are represented as follows (using GFs for composite verbs):

$$\begin{bmatrix} \textsc{clause\_DepRel} & \text{discourse} \\ \textsc{clause\_POS} & \text{INTJ} \\ \textsc{disc\_discourse\_DepRel} & \text{discourse} \\ \textsc{disc\_discourse\_POS} & \text{INTJ} \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} \textsc{clause\_DepRel} & \text{root} \\ \textsc{clause\_POS} & \text{VERB} \\ \textsc{clause\_prec} & , \\ \textsc{clause\_succ} & . \\ \textsc{NP\_nsubj\_Case} & \text{Nom} \\ \textsc{NP\_nsubj\_DepRel} & \text{nsubj} \\ \textsc{NP\_nsubj\_Number} & \text{Sing} \\ \textsc{NP\_nsubj\_Person} & 1 \\ \textsc{NP\_nsubj\_POS} & \text{PRON} \\ \textsc{NP\_nsubj\_PronType} & \text{Prs} \\ \textsc{verb\_Aspect} & \text{Imp} \\ \textsc{verb\_Mood} & \text{Ind} \\ \textsc{verb\_Tense} & \text{Fut} \\ \textsc{verb\_VerbForm} & \text{Fin} \\ \textsc{verb\_Voice} & \text{Act} \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} \textsc{clause\_DepRel} & \text{xcomp} \\ \textsc{clause\_POS} & \text{VERB} \\ \textsc{NP\_obj\_Definite} & \text{Ind} \\ \textsc{NP\_obj\_DepRel} & \text{obj} \\ \textsc{NP\_obj\_Number} & \text{Sing} \\ \textsc{NP\_obj\_POS} & \text{NOUN} \\ \textsc{NP\_obj\_PronType} & \text{Art} \\ \textsc{verb\_VerbForm} & \text{Inf} \end{bmatrix} \tag{9}$$

The prefix `clause`, `NP`, `verb` or `disc` corresponds to the syntactic unit as described in Sections 3.1–3.4.

B Decision Tree

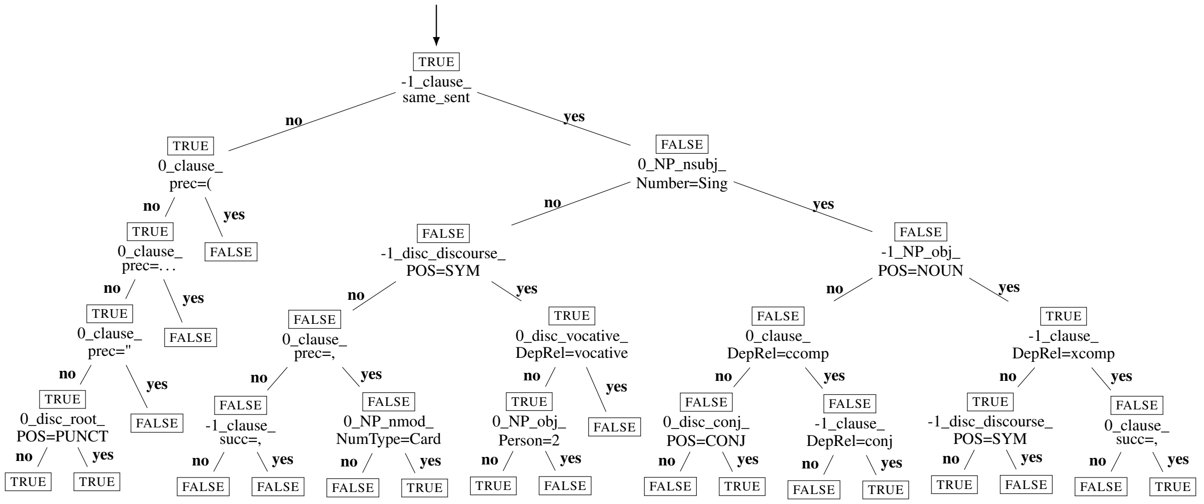

Figure 3: Decision tree learned on `eng.sdrt.stac` (using GFSs). "TRUE" are segment starts. The number prefixed to a feature is the offset to the current clause, e.g. feature of the preceding clause start with "-1".