


ARTICLE

Toward a shallow discourse parser for Turkish

Ferhat Kutlu¹ , Deniz Zeyrek¹  and Murathan Kurfalı² 

¹Graduate School of Informatics, Cognitive Science Department, Middle East Technical University, Çankaya/Ankara, Turkey and ²Linguistics Department, Stockholm University, Stockholm, Sweden

Corresponding author: Ferhat Kutlu; Email: ferhat.kutlu@metu.edu.tr

(Received 5 April 2021; revised 21 May 2023; accepted 19 June 2023)

Abstract

One of the most interesting aspects of natural language is how texts cohere, which involves the pragmatic or semantic relations that hold between clauses (addition, cause-effect, conditional, similarity), referred to as discourse relations. A focus on the identification and classification of discourse relations appears as an imperative challenge to be resolved to support tasks such as text summarization, dialogue systems, and machine translation that need information above the clause level. Despite the recent interest in discourse relations in well-known languages such as English, data and experiments are still needed for typologically different and less-resourced languages. We report the most comprehensive investigation of shallow discourse parsing in Turkish, focusing on two main sub-tasks: identification of discourse relation realization types and the sense classification of explicit and implicit relations. The work is based on the approach of fine-tuning a pre-trained language model (BERT) as an encoder and classifying the encoded data with neural network-based classifiers. We firstly identify the discourse relation realization type that holds in a given text, if there is any. Then, we move on to the sense classification of the identified explicit and implicit relations. In addition to in-domain experiments on a held-out test set from the Turkish Discourse Bank (TDB 1.2), we also report the out-domain performance of our models in order to evaluate its generalization abilities, using the Turkish part of the TED Multilingual Discourse Bank. Finally, we explore the effect of multilingual data aggregation on the classification of relation realization type through a cross-lingual experiment. The results suggest that our models perform relatively well despite the limited size of the TDB 1.2 and that there are language-specific aspects of detecting the types of discourse relation realization. We believe that the findings are important both in providing insights regarding the performance of the modern language models in a typologically different language and in the low-resource scenario, given that the TDB 1.2 is 1/20th of the Penn Discourse TreeBank in terms of the number of total relations.

Keywords: Discourse relation; Classification; Pre-trained language model; Encoding; Cross-lingual transfer learning

1. Introduction

Turkish is a language of more than 80 M speakers and belongs to the Turkic sub-family of the Altaic language family. It has a complex morphology, where suffixation is a major tool of both derivation and inflection and hence poses several challenges for NLP (Ofłazer and Saraçlar 2018). Despite its large number of speakers, its entrance to the NLP field is rather recent, and language technology tools have been attempted only in the last few decades (Ofłazer and Bozşahin 1994). The interest in Turkish NLP and language technology tools have been increasing with interest in sentence-level tasks such as named entity recognition (Seker and Eryiğit 2017; Akkaya and Can 2021) as well as semantics (Eryiğit, Nivre, and Ofłazer 2008; Çakıcı, Steedman, and Bozşahin 2018).



Understanding natural language texts not only requires the knowledge of sentence structure and sentence meaning but also the knowledge of how texts cohere. Although naturally easy for human language users, understanding how texts cohere is still a challenge for Natural Language Understanding (NLU) because the task requires to go beyond words and clauses. Recently, to enable research on linguistic structures above the sentence level and to enhance language technology applications that exploit such structures (such as text summarization, dialogue systems, information retrieval, and machine translation), there have been rigorous attempts to create linguistic corpora annotated for semantics or discourse, for example Framenet (Baker, Fillmore, and Lowe 1998), Propbank (Palmer, Gildea, and Kingsbury 2005), Groningen Meaning Bank (Bos 2013), and the Penn Discourse TreeBank (PDTB) (Prasad *et al.* 2008). To date, the largest annotated discourse corpus for English is the PDTB containing over 40,600 discourse-level annotations on Wall Street Journal texts (Prasad, Webber, and Joshi 2014). It annotates one of the building blocks of discourse structure, that is, discourse relations (DRs), which are pragmatic or semantic relations that hold between clauses (addition, cause-effect, conditional, similarity).

Discourse relations may be realized explicitly or implicitly. In the lexicalized approach of the PDTB framework, connectives are considered as lexico-syntactic devices that signal the presence of a discourse relation. Relations that are made salient by discourse connectives are referred to as explicit relations. Discourse relations may also be instantiated without any discourse connective, known as implicit relations. Even in these cases, the semantic relation between text segments (referred to as the arguments of a relation) can be easily inferred by humans. The PDTB treats discourse connectives as discourse-level predicates that take two abstract objects as arguments (events, states, and propositions Asher 1993) and annotates explicitly and implicitly conveyed relations, their arguments, and senses. This annotation style has triggered an active line of research in discourse parsing, particularly in English. However, many languages still lag behind such developments presenting a challenge to universal end-to-end NLU pipelines.

The present paper aims to move toward filling this gap by working on Turkish discourse relations in corpora annotated in the PDTB style, the Turkish Discourse Bank (TDB) 1.2 and the Turkish subpart of the TED Multilingual Discourse Bank (T-TED-MDB). We present the results of our ongoing work toward an end-to-end discourse parser for Turkish. The task of discourse parsing aims to uncover all the underlying discourse relations, along with their arguments and senses, if any, in a given text. It involves various sub-tasks, each targeting different components of discourse relations (see Section 2.2). Our current pipeline sidesteps the problem of argument span extraction by performing DR realization type identification directly on texts. Specifically, we train models to perform two tasks: (i) DR realization type identification (explicit, implicit, etc.) and (ii) classifying the Level-1 senses of explicit and implicit relations. Classifying the sense of implicit relations is one of the most challenging task in discourse parsing. We model both tasks as multi-class classification tasks and present a series of experiments that use the modern pre-trained language models (PLM) and neural network-based classifiers. The main contributions of the present work are summarized below and depicted in Fig. 1:

- We perform the most thorough analysis on discourse parsing on Turkish. Specifically, a DR realization type classifier and Level-1 sense classifiers are built for explicit and implicit discourse relations.
- To circumvent the training data scarcity problem that arises due to the cost of annotation, we run two kinds of experiments limiting ourselves to DR realization type classification: (a) First, we investigate the effect of adding an additional language to BERT by preparing a custom multilingual dataset consisting of two languages (English and Turkish) and run experiments with multilingual pre-trained language models. (b) Secondly, the efficiency of cross-lingual transfer learning techniques is investigated through multilingual language models, and the results of our experiments over the monolingual (Turkish) PLM and the multilingual PLM are compared.

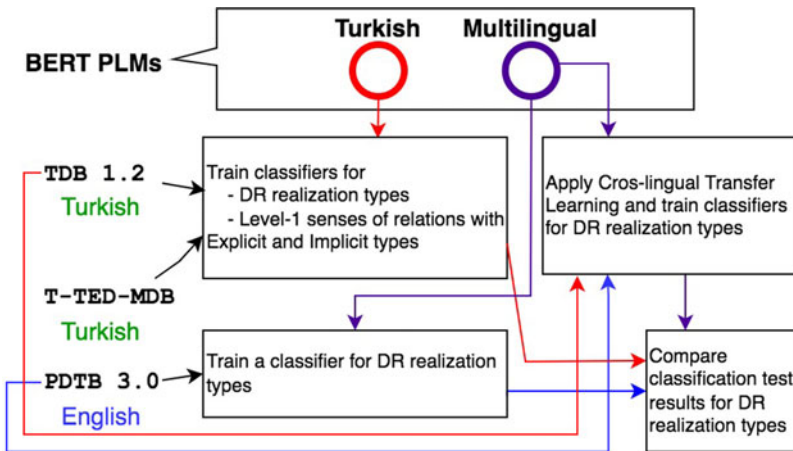


Figure 1. Symbolic representation of our approach

- Thus, by aggregating the TDB 1.2 with the PDTB 3.0, to the best of our knowledge, we perform the first cross-lingual investigation on the DR realization type identification task.
- In order to evaluate the generalization abilities of the DR realization type classification model, we test and report its out-of-domain performance using the Turkish part of the TED Multilingual Discourse Corpus.

The rest of the paper proceeds as follows. Section 2 reveals the PDTB annotation principles with a focus on DR realization types and then describes the sub-tasks of our shallow discourse parsing pipeline undertaken in the current research. Section 3 provides the literature review on discourse parsing as well as the necessary background to understand the models and techniques employed in our experiments. Section 4 overviews the datasets used in the experiments along with the details of the experimental setup in order to facilitate the replicability of our results. Section 5 presents the results of our experiments and discusses their implications in detail. Finally, Section 6 concludes the paper with final thoughts and further suggestions on future directions.

2. Shallow discourse parsing

Shallow discourse parsing refers to uncovering local discourse relations in a text as they are defined according to the PDTB. Since the PDTB annotation scheme was used for all versions of the TDB, in this section, we briefly introduce the PDTB annotation principles, focusing on DR realization types. Then, we describe the sub-tasks of the shallow discourse parsing pipeline undertaken in the current research.

2.1 Discourse relation realization types

As already mentioned, the PDTB mainly annotates explicitly and implicitly conveyed relations together with their arguments (Arg1, Arg2) and senses. This section overviews the discourse relation realization types in the PDTB 3.0 (Webber, Prasad, and Lee 2019), which extends the earlier version keeping the rules and principles of the PDTB framework intact. The sense hierarchy is also revised. In the examples, discourse connectives, where available, are underlined for clarity.

Explicit DRs are those relations that are explicitly signaled through lexico-syntactic elements. They may be realized across or within sentences and involve coordinating conjunctions (*and*, *so*,

but) subordinating conjunctions (*because, after, while*), adverbials (*however, additionally, consequently*), and prepositional phrases (*in summary, on the contrary*). These markers are referred to as explicit discourse connectives, as shown in example 1. In the PDTB 3.0, explicit discourse connectives have been extended to cover subordinators (i.e., prepositions such as *because of, by, despite, for, from*) to more fully annotate intra-sentential relations and are only annotated when they take clausal complements, as in example 2.

- (1) The city's Campaign Finance Board has refused to pay Mr. Dinkins \$95,142 in matching funds because his campaign records are incomplete.
- (2) Eliminate arbitrage and liquidity will decline instead of rising, creating more volatility instead of less.

Implicit DRs: In the absence of a connective supporting a discourse relation, readers can infer the meaning of the relation, and the annotators are asked to insert an explicit marker that best conveys the sense of the inferred relation. These are called implicit connectives. While the PDTB 2.0 only annotated implicit DRs that hold between sentences (example 3), the PDTB 3.0 annotated implicit relations within sentences as well (e.g., between VPs or clauses conjoined implicitly by punctuation, as in example 4) (Zhao and Webber 2021). The PDTB 3.0 also annotates implicit relations when they co-occur with explicit relations, as will be exemplified in examples 12 and 13 below.

- (3) So much of the stuff poured into its Austin, Texas, offices that its mail rooms there simply stopped delivering it. (Implicit = so) Now, thousands of mailers, catalogs, and sales pitches go straight into the trash.
- (4) Father McKenna moves through the house praying in Latin, (Implicit = and) urging the demon to split.

Alternative Lexicalization (AltLex): When an implicit discourse relation is inferred to hold between or within sentences but the insertion of an implicit connective in the relation is perceived redundant, the relation is referred to as being alternatively lexicalized. Expressions that are inferred to confirm the presence of a discourse relation are annotated as AltLex (example 5). The PDTB 3.0 extended AltLexes to cover AltLexC relations, that is, lexico-syntactic constructions that unambiguously signal discourse relations, such as example 6.

- (5) After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%. The Reason: Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold.
- (6) Crucial as these elections are for Greece, pressing issues of state are getting lost in the shuffle.

Entity Relations (EntRels): Where no discourse relation can be inferred between adjacent sentences and adjacent sentences form entity-based coherence with the same entity being realized in both sentences either directly or indirectly, the discourse relation is annotated as an EntRel (example 7).

- (7) Pierre Vinken, 61 years old, will join the board as a non-executive director Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

No Relations (NoRels): Where a semantic relation between adjacent sentences cannot be identified, the adjacent pair of sentences are annotated as NoRel (example 8).

- (8) Mr. Rapanelli met in August with U.S. Assistant Treasury Secretary David Mulford. Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks.

Hypophora: Where a question is asked and a meaningful response is provided, Hypophora is annotated. This is a new inter-sentential relation in the PDTB 3.0, and it stands for discourse relations involving dialogue acts, which cannot be instantiated through connectives (example 9).

- (9) But can Mr. Hahn carry it off? In this instance, industry observers say, he is entering uncharted waters.

In summary, in the PDTB 3.0, not only explicit and AltLex discourse relations but also implicit discourse relations are annotated both across and within sentences and are assigned a sense tag from the revised hierarchy of senses. The PDTB 3.0 keeps the four Level-1 senses of earlier versions, namely Expansion, Comparison, Contingency, and Temporal. These are refined to more specific senses at the second level. For symmetric relations, the sense hierarchy stops at the second level; for asymmetric relations, the hierarchy provides third-level senses to define the semantic contribution of each argument.^a (The full sense hierarchy is provided in Appendix A). EntRels, NoRels, and Hypophora are annotated as inter-sentential relations and not assigned a sense tag.

Multiple Relations: In addition to the DRs overviewed so far, the PDTB 3.0 annotates multiple relations in certain cases. Multiple relations are those relations that can hold between a pair of spans. For example, when there are co-occurring explicit connectives between the same spans, multiple relations are annotated by creating two tokens (examples 10, 11).

- (10) Small businesses say a recent trend is like a dream come true: more-affordable rates for employee-health insurance, initially at least. But then they wake up to a nightmare.
- (11) Small businesses say a recent trend is like a dream come true: more-affordable rates for employee-health insurance, initially at least. But then they wake up to a nightmare.

Also, in cases where annotators infer a sense separate from the one conveyed by an explicit connective or AltLex or AltLexC, separate implicit relations are annotated by inserting implicit connectives (examples 12, 13).

- (12) We've got to get out of the Detroit mentality and be part of the world mentality, declares Charles M. Jordan, GM's vice president for design. . .
- (13) We've got to get out of the Detroit mentality and (Implicit=instead) be part of the world mentality, declares Charles M. Jordan, GM's vice president for design. . .

2.2 Shallow discourse parsing sub-tasks

A typical shallow discourse parser consists of three main sub-tasks: (i) connective identification, (ii) argument extraction, and (iii) sense classification. In the current paper, we have adopted a different approach and merged and converted the first two tasks into the identification task of a

^aSymmetric relations are those where (Arg1, Arg2) and (Arg2, Arg1) are semantically equivalent. A relation that is not symmetric is defined as asymmetric (Webber *et al.* 2016).

discourse relation in a given text piece, referred to as *DR realization type identification*. That is, our approach does not perform distinct connective and argument extraction, but it is still able to identify the discourse relation, if there is any, in a given text piece and further disambiguates the discourse relations in terms of the senses. Yet, we would like to note that the model is trained on both inter- and intra-relations without the knowledge of distinction^b and is unable to detect multiple discourse relations simultaneously. If the given text piece has more than one different DRs, the model will be able to find only one of them.

DR realization type identification: This is one of the least studied aspects of discourse. Here, it is modeled as a six-way classification task which aims to identify the specific realization type of the discourse relation in the given text. Although the usage disambiguation of discourse connectives (distinguishing between the connectives' discourse and non-discourse role) has been investigated for many languages as discussed in Section 3, and implicit relation identification has also been targeted, these have often been considered as tasks on their own. The multi-level identification of relation realization type is a challenging task, which to our knowledge has not been tackled before.

Sense classification of discourse relations: This is the most popular sub-task of discourse parsing, where the aim is to find the sense conveyed by a given explicit or implicit discourse relation. The implicit classification task is rather (in)famous for its challenging nature due to the lack of an explicit signal, that is, the discourse connective, and renders the task extremely challenging, where the models must encode the semantics of each argument correctly. Due to its challenging nature, the task is most commonly limited to only the Level-1 senses in the PDTB 3.0 sense hierarchy; hence, in the current work, it is modeled as a four-way classification task. To the best of our knowledge, the four-way sense classification of implicit discourse relation is the first attempt over Turkish data. The second and third-level senses are out of scope of the current work.

3. A brief survey of related work

3.1 Attempts toward the development of discourse parsers

Prior to the work of Lin, Ng, and Kan (2014), research only targeted individual tasks involved in discourse parsing. For example, the task of automatically identifying explicit discourse connectives was shown to be doable with high accuracy with the help of linguistic features and information about the connective's surrounding context. In an early paper, Pitler *et al.* (2008) developed decision trees (i) to distinguish between explicit and implicit discourse relations in the PDTB 2.0, (ii) to distinguish tokens of each Level-1 sense relation from all the others, (iii) to carry out four-way sense classification of all tokens and a separate four-way sense classification of tokens with explicit connectives. Their model reached a higher performance in the classification of explicit discourse connectives than implicit ones. One of the earliest studies on the identification of discourse *versus* non-discourse usage of explicit connectives has been carried out by Pitler and Nenkova (2009) over the PDTB 2.0. Feeding syntactic features extracted from the arguments of discourse connectives into a maximum entropy classifier,^c the authors reached an F-Score of 0.92 in explicit discourse connective disambiguation and 94% accuracy in the four-way sense classification of explicit discourse connectives.

Later work has reached highly successful results in domain-specific applications. Ramesh *et al.* (2012) used various supervised machine-learning-based algorithms for automatically identifying explicit discourse connectives in the BioDRB corpus and proposed a hybrid classifier based

^bThis means that at the time of inference, the model considers the possibility of an intra-sentential relation being a NoRel although such a combination is not possible. Augmenting intra-/inter-sentential information to the training procedure would probably improve the training, but since such information is not always available, we wanted to work with what we can safely assume to exist in the corpora.

^c<https://github.com/mimno/Mallet>

on a Conditional Random Fields-based classifier and a combination of instance pruning, feature augmentation, and domain adaptation techniques. Extracting syntactic features such as the part-of-speech (POS) tags of the tokens, the syntactic labels of the immediate parent of the token's POS in the parse tree, and the POS tags of the left sibling (the token to the left of the current word inside the innermost constituent), an F-Score of 0.76 is reached. Gopalan and Devi (2016) used fewer linguistic features relevant to discourse and employed machine-learning models to automatically extract explicit discourse connectives, their senses, and arguments. They reported an F-Score of 0.85 in the classification of explicit discourse connectives' senses with the Conditional Random Fields algorithm.

Work on non-English languages such as Arabic discovered features that contribute to the disambiguation of the discourse usage of explicit connectives. Alsaif and Markert (2011) worked on an Arabic corpus where explicit discourse connectives are marked. They used syntactic features such as the position of the potential connective (sentence-initial, sentence-medial, or sentence-final), lexical features of the surrounding words, the POS tags of the words, and the syntactic category of the parent of the potential connective. The authors also discovered the predictive role of the infinitive form of the verb in the second argument of prepositional connectives and achieved an F-Score of 0.78 in explicit discourse connective recognition with the best feature combination.

A recent work by Başbüyük and Zeyrek (2023) disambiguates the discourse usage of Turkish connectives with a rule-based and machine-learning approach using different sets of linguistic rules for discourse connectives belonging to different syntactic classes. The machine-learning approach achieves an F1-score of 84.44 for single/phrasal connectives with the Fast Forest Binary algorithm and an F1-score of 80 for suffixal connectives with the Averaged Perceptron Binary algorithm with the best feature combination.

While it could be said that explicit discourse relation recognition is largely a solved problem, especially for well-studied languages such as English, identification of implicit relations still remains a challenge. Marcu and Echiabi (2002) have been the first to identify implicit relations by removing discourse connectives to cheaply gather large amounts of training data. They also discovered that word pairs are indicative of implicit relations (e.g., the pair *embargo . . . legally* was a good indicator of contrast relations) and used them in extracting large amounts of data. They reached high accuracies with this technique of obtaining artificial implicit relations. Particularly, two of their classifiers were successful in distinguishing between Cause-Explanation-Evidence versus Elaboration (93%) and Cause-Explanation-Evidence versus Contrast (87.3%).

A later paper by Pitler, Louis, and Nenkova (2009) predicted the Level-1 senses of implicit relations in the PDTB 2.0 in a realistic setting taking advantage of linguistically informed features as well as lexical pairs from unannotated text. They achieved an F-Score of 0.76 with the best combination of their features (polarity+Inquirer tags+context). Lin, Kan, and Ng *et al.* (2009) took parser production rules as the main source of features and worked on the Level-2 senses of the PDTB 2.0 hierarchy. They showed that syntactic patterns could contribute to predicting the senses of implicit relations. In CoNLL-2015 Shared Task, the feature-based work of Rutherford and Xue (2014) presented statistical classifiers for identifying the senses of implicit relations. The authors introduced novel feature sets that exploit distributional similarity and coreference information. They showed that Brown cluster pairs (Brown *et al.* 1992) work well in implicit relation recognition.

A breakthrough is seen in the field with the development of an end-to-end discourse parser (for English) by Lin *et al.* (2014). Working on the PDTB 2.0, the authors produced a PDTB-styled parser, constructing a thorough pipeline for parsing the text in seven sequential steps titled as connective classification, argument labeling, argument position classification, argument extraction, explicit relation classification, and sense recognition of explicit as well as non-explicit relations. The authors classified and labeled discourse relations and the attribution spans, where relevant. Their parser can parse any unrestricted English text into its discourse structure in the PDTB style.

The best F1-Scores are reported as 0.87 for the explicit classifier, and 0.4 for the non-explicit classifier.

Lately, shallow discourse parsing has been attempted in a series of shared tasks. As in the work of Mukherjee *et al.* (2015), shallow discourse parsing is mostly conducted by using syntactic and semantic features for classification. It is also attempted in CoNLL Shared Task (2015), to which 16 teams participated by using a piece of newswire text as input and returning relations in the form of a discourse connective (either explicit or implicit) with two arguments. Each team developed an end-to-end system that could be regarded as variations of Lin *et al.* (2014), detecting and categorizing individual discourse relations and returning a set of relations contained in the text. The best system achieved an F1-Score of 0.24 on the blind test set, reflecting the serious error propagation problem in such a system (Xue *et al.* 2015, 14).

The shared task DISRPT (2019) (DR Parsing and Treebanking) was held on discourse unit segmentation across formalisms, including shallow discourse parsing, aiming to promote the convergence of resources and a joint evaluation of discourse parsing approaches. The corpora included 15 datasets in 10 languages, 12 of which target elementary discourse unit segmentation, and three dedicated to explicit connective annotation (Zeldes *et al.* 2019). In the overall evaluation, ToNy (Muller, Braud, and Morey 2019) performed the best on most of the datasets reaching an F-Score of 0.9 in the average of all its tests. Turkish is also represented in the dataset with the TDB 1.0, where only explicit discourse connectives are annotated. On this data, ToNy (Muller *et al.* 2019) obtained the best results in discourse connective detection on plain, unannotated data, reaching an F-Score of 0.85. In the DISRPT (2021) shared task event, the approach of the group DiscoDisco increased the F1-Score of Turkish explicit discourse connective detection sub-task to 0.94.

One of the latest noteworthy achievements has been reported by Liang, Zhao, and Webber (2020). The authors worked on the PDTB 3.0, where implicit relations are annotated both at the inter-sentential and intra-sentential levels. In addition to these stand-alone implicits, implicit sense relations between the arguments of explicit relations are also annotated. In a series of experiments, the authors first recognized the location of implicits, then they recognized their senses, arguing that the data annotated in this way simplifies the difficult problem of sense-labeling of implicits.

To sum up, the supervised classification algorithms described above have been able to classify discourse relations, particularly explicit ones, very successfully in texts. But the results lack general impact, as less-studied, low-resource languages have hardly been targeted. Moreover, sense-labeling of implicit discourse relations still lags behind that of explicit relations. The next section deals with the impact of deep learning models in the field of discourse understanding as background to the experiments conducted in the current work.

3.2 Pre-trained language models: The paradigm shift in NLP domain

The advent of attention mechanism has been a breakthrough in NLP (Bahdanau, Kyunghyun, and Yoshua 2015; Luong, Pham, and Manning 2015) which, consequently, gave rise to the transformer architecture (Vaswani *et al.* 2017). BERT is undoubtedly the most famous language model that is based on the transformer architecture (Devlin *et al.* 2019). BERT uses a masked language model which randomly masks 15% of the tokens from the input in order to predict the original vocabulary by feeding the final hidden vectors (corresponding to the masked tokens) into an output softmax over the vocabulary. Discourse relation classification benefited from BERT. For example, Nie, Bennett, and Goodman (2019) showed that BERT can outperform previous state-of-the-art models in implicit discourse relation classification, and Kishimoto, Murawaki, and Kurohashi (2020) adapt BERT to the task, managing to reach an F1-Score of 0.59.

3.3 Cross-lingual transfer learning

In recent years, cross-lingual transfer has become an active area of research in favor of low-resource languages. It is a fairly specific way of training models using the data available for languages with ample resources so that it can solve the same task in the target low-resource language(s). Cross-lingual transfer learning serves many research domains including the construction of bilingual dictionaries (Wang *et al.* 2019), zero-shot translation (Johnson *et al.* 2017), spoken language understanding (Yazdani and Henderson 2015), semantic utterance classification (Dauphin *et al.* 2014), entity extraction from Web pages (Pasupat and Liang 2014), fine-grained named entity typing (Ma, Cambria, and Gao 2016), cross-lingual document retrieval (Funaki and Nakayama 2015), relation extraction (Levy *et al.* 2017), multilingual task-oriented dialog (Schuster *et al.* 2019), and event detection (Caselli and Üstün 2019).

A more specific case of transfer learning is known as *zero-shot learning*, where the classifier is able to classify examples it is never exposed to during the training. In the cross-lingual scenario, this often translates into training the classifier in one language and applying it to other languages with a minimal performance loss. The advantage of cross-lingual transfer learning, in general, is being able to leverage information from high-resource languages into the low-resource ones. In the case of discourse parsing, it is highly relevant as almost all of the non-English languages lack large manually annotated datasets.

4. An overview of the data, experiments, and the experimental setup

In the current paper, BERT is used in order to learn the labels over discourse relations, taking advantage of its capacity to obtain the intrinsic representation of each word while preserving the linguistic properties of the text. This section details the datasets and overviews the experimental setup of the experiments described in Section 5.

4.1 Datasets

The datasets used in the paper involve the following: (i) The TDB 1.2, (ii) the PDTB 3.0, and (iii) the Turkish section of the TED-MDB.

4.1.1 Turkish Discourse Bank

The TDB 1.2 is chosen as the main data source. It is a 40,000-word discourse-level resource of Turkish created by manually annotating a corpus of modern Turkish texts written between 1990 and 2000 (Zeyrek and Webber 2008). The annotation principles of the TDB 1.2 and the annotated categories closely follow those of the PDTB; most notably, all discourse relation realization types described in Section 2.1 are spotted and annotated together with their binary arguments and senses, where relevant.

The basic principle in annotating explicitly marked relations is the PDTB's minimality principle, that is, the annotators are asked to select the shortest text spans (e.g., clauses or sentences) that are necessary and sufficient to interpret a discourse relation encoded by a connective. The datasets over which the two corpora are built differ; while the PDTB is built over Wall Street Journal Corpus, the TDB is built over multiple genres (newspaper editorials, fiction, popular magazines). Other differences involve (i) the way different types of discourse connectives are annotated (e.g., suffixal connectives are annotated as a type of explicit connectives (Zeyrek and Basıbuyuk, 2019)) (ii) a small number of new Level-2 sense tags are spotted and annotated (Zeyrek and Kurfalı 2017), and (iii) attribution and AltLexC relations have not been annotated in the TDB 1.2 so far.

In the annotation stage, any possible discourse relation is searched and annotated by going through the texts sentence by sentence, similar to the incremental processing of discourse.

Table 1. Distribution of DR types and Level-1 senses in the TDB 1.2 and the PDTB 3.0

DR type	TDB	PDTB	Level-1 Sense	TDB	PDTB
Explicit	1467 (38.2%)	24,240 (45.3%)	Comparison	448 (12.9%)	8,399 (18.25%)
Implicit	1743 (44.9%)	21,782 (40.7%)	Contingency	702 (20.3%)	11,503 (24.99%)
AltLex	146 (3.8%)	1498 (2.8%)	Expansion	1,700 (49%)	20,266 (44.04%)
Hypophora	77 (2%)	146 (0.27 %)	Temporal	617 (17.8%)	5,854 (12.72%)
EntRel	233 (5.9%)	5538 (10.3 %)			
NoRel	203 (5.1%)	287 (0.5 %)			

Annotations were performed by a team of trained graduate students, sustaining a minimum value of $\kappa = 0.7$ for inter-annotator agreement (Zeyrek and Kurfalı 2017). Even though this is below the normal threshold of 0.8, due to the ambiguity of coherence relations, the annotation task is very hard, forcing the team to take 0.7 as a satisfactory level as suggested by Spooren and Degand (2010).

The TDB 1.2 contains 3987 discourse relations (Table 1), where discourse senses are annotated on the basis of the PDTB 3.0 sense tag-set (Appendix A) (Zeyrek and Er 2022). Furthermore, explicit and implicit relations and AltLexes are annotated both at the inter-sentential and intra-sentential level. EntRels, NoRels, and Hypophora (13% of the corpus) are annotated only at the inter-sentential level.

4.1.2 Penn Discourse Treebank 3.0

The Penn Discourse Treebank is the largest text corpus annotated for discourse relations. The latest release, the PDTB 3.0, is an enriched version of the PDTB 2.0 with the addition of 13K new annotations, which are mainly intra-sentential relations that were not annotated in the previous edition (Webber *et al.* 2019). The PDTB 3.0 includes a total of 53631 annotations, the distribution of which is provided in Table 1.

4.1.3 TED-Multilingual Discourse Bank

The TED-MDB is a multilingual corpus that follows the same annotation principles of the PDTB and the TDB (Zeyrek, Mendes, and Kurfalı 2018, Zeyrek *et al.* 2020) follows the PDTB 3.0 sense hierarchy. Yet, unlike those corpora, the TED-MDB is annotated on the subtitles of six TED talks. In total, there are 695 discourse relations (317 explicit and 210 implicit relations) in the Turkish part of the corpus.

4.2 Experimental setup

Each task is modeled as a multi-way classification task. As we do not assume a separate identification of arguments, each discourse relation is presented as one unit to BERT; hence, in practice, the models do not have any information regarding the boundaries of the discourse arguments but see them as one continuous text piece.

For each dataset, 10% of the data is allocated as the validation set and another 10% as the test set. The distribution of labels in each set is provided in Table 2. For each task, we fine-tune BERT following the standard practice suggested by Devlin *et al.* (2019).

Following the previous work, we set the maximum sequence length to 128. We use AdamW optimizer (Loshchilov and Hutter 2017) with the learning rate of $5e - 5$. We also apply a learning rate warm-up where the learning rate is linearly increased from 0 to $5e - 5$ over the first 10%

Table 2. Summary of the train, development, and test splits of the TDB 1.2 used in the experiments. The upper part presents the distribution of labels in the DR realization type classification experiments and the lower part presents the sense experiments over explicit and implicit DRs.

TYPE	Train	Dev	Test
AltLex	113	13	20
EntRel	196	24	13
Explicit	1157	147	163
Hypophora	68	2	7
Implicit	1394	186	163
NoRel	167	15	21

Level-1 Sense	Explicit			Implicit		
	Train	Dev	Test	Train	Dev	Test
Comparison	205	23	31	130	15	17
Contingency	206	29	33	264	35	34
Expansion	429	57	54	879	118	93
Temporal	317	38	45	121	18	19

of iterations and then linearly decreases to 0. The models are fine-tuned for 50 epochs, with the batch sizes of 16. We apply early stopping with patience^d of 25 according to the performance on the development set. The models are evaluated four times in each epoch and the one with the best development performance is stored as the final model. As the target metric, we use Macro-F1. Yet, for each experiment, we report the performance of our models on the set via accuracy, recall, precision, Macro-F1, and Weighted F1,^e calculated as:

$$Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP} \quad F1 = \frac{2 * Recall * Precision}{Precision + Recall} \quad (1a)$$

$$Macro-F1 = \frac{1}{N} \sum_i^N F1_i \quad (where N is the number of labels) \quad (1b)$$

$$Weighted-F1 = \sum_i^N w_i \cdot F1_i \quad where(w_i)is proportional to the frequency of class i \quad (1c)$$

All the training is performed on a single T4 GPU. The experiments for the development of a classification model gave rise to the *Bert_MultiClass* model, built based on the *transformers.TFBertModel*^f class (More details about the classification model, the method of fine-tuning, and the tests are provided in Appendix B).

^dEarly stopping is a feature that enables the training to be automatically stopped when a chosen metric has stopped improving. Patience is the number of epochs without improvement, after which training will be early stopped.

^ehttps://scikit-learn.org/stable/modules/model_evaluation.html

^fhttps://huggingface.co/docs/transformers/v4.15.0/en/model_doc/bert#transformers.TFBertModel

Table 3. An overview of the conducted experiments

Experiment #	Encoding PLM	Dataset	Purpose
5.1	Turkish BERT	TDB 1.2 T-TED-MDB	To explore the effectiveness of a monolingual BERT for both in-domain and out-domain datasets in the classification of DR realization types and Level-1 senses
5.2	Turkish BERT Multilingual BERT	TDB 1.2 PDTB 3.0	To explore the effectiveness of multilingual data aggregation on DR realization type classification

5. Results and discussion

In this section, we report and discuss the results of our experiments on Turkish discourse parsing. We carried out two sets of experiments. The first set involves in-domain experiments on a held-out test set from the TDB 1.2, where we identify the realization type of the discourse relation that holds in a given text span, if there is any (Section 5.1.1). Then, we move on to the sense classification of the identified explicit and implicit relations (Section 5.1.2). We report the out-domain performance of our models in order to evaluate their generalization abilities, using the T-TED-MDB (Section 5.1.3). In the second set of experiments, through a series of cross-lingual transfer learning experiments, we explore the effect of multilingual data aggregation on DR realization type classification (Section 5.2) and discuss the extent to which the lack of training data that arises due to the cost of manual annotation can be alleviated in a low-resource scenario such as ours. Table 3 provides an overview of the conducted experiments.

5.1 Monolingual experiments

The first set of experiments aims to explore the effectiveness of a monolingual BERT PLM for the target tasks. Despite having quickly become the de facto way of performing NLP tasks in recent years, such pre-trained models have seen very limited attention for non-English languages, where Turkish is no exception. We address this shortcoming and investigate the performance of a Turkish BERT model using the Hugging Face library⁸ with the available annotated data in Turkish. In what follows, we present the results of DR realization type classification for the first time.

5.1.1 DR realization type classification

As defined in Section 2.2, DR realization type classification focuses on identifying how precisely discourse relations are realized in a given text span, given the PDTB's five relation realization types (implicit, explicit, AltLex, Hypophora, EntRel). If no such relation is found, the model is supposed to label the text as having a NoRel relation, mimicking the PDTB annotation style. The results of our experiments are provided in Table 4.

According to the results, the Turkish BERT model achieves almost 74% accuracy and a Macro-F1 Score of 0.58 over all relations. The relatively low Macro-F1 Score suggests that the task is more challenging than it looks; however, it must be highlighted that the model does not have access to

⁸<https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased>

Table 4. DR realization type classification results over the TDB 1.2

DR Type	F1-score	Recall	Precision
AltLex	0.63	0.58	0.69
EntRel	0.32	0.40	0.32
Explicit	0.90	0.89	0.92
Hypophora	0.74	0.79	0.70
Implicit	0.77	0.78	0.76
NoRel	<u>0.11</u>	0.11	0.12
Accuracy	73.90%		
Macro-F1	0.58		
Weighted F1	0.77		

any information regarding the argument boundaries or whether there is a connective in the text span or not. Hence, we find the results to be promising.

Of all relation realization types, the explicitly realized ones turn out to be the most easily identifiable relations with the F1-Score of 0.9. This result suggests that the model implicitly learns to recognize discourse connectives. On the other end of the spectrum lie the EntRels and NoRels which are identified very poorly. Hypophora relations, which are the least frequent relations in the data, are surprisingly identified well.

The model manages to identify Hypophora relations almost as good as implicits. Although it is hard to draw any definite conclusions due to the limited sample size, considering that Hypophora relations consist of a rather fixed template, a question-answer pair where commonly the first argument expresses a full question and the other provides an answer, it is plausible that the model learns to associate this form with the Hypophora label.^h Yet, we would like to emphasize that at this point any observation regarding the infrequent relation types, especially EntRel and Hypophora relations, are only informed guesses due to their limited size and must be justified with further experiments with more data.

In order to gain further insight into the model's decisions, we first analyzed the results in the form of a confusion matrix provided in Appendix C, Fig. C1. According to the confusion matrix, the model frequently mixes EntRels with implicits. However, confusing EntRels with implicits is not really unexpected: Zeyrek and Kurfalı (2017) report that human annotators also struggle with telling these two relations apart. Implicit relations conveying an Expansion sense (especially the Level-2 sense of "level-of-detail") look very similar to EntRels as they also tend to talk about a common entity.

As for NoRels, they are almost always classified as an implicit relation (18 out of 21 relations). Therefore, it is clear that the model did not learn the difference between these two relations. This is probably the case because the number of NoRels is pretty limited in the data, and more importantly, these non-relations do not have any specific clue for the model to pick up (e.g., a set of recurrent tokens such as connectives); so the model tends to classify them as the more frequent implicit relations that also do not have any characteristic explicit clue.

Next, we extracted the sentence embeddings of the textual elements formed by concatenating Arg1, Arg2 and the discourse connective (if available) of each discourse relation in the TDB 1.2. We computed the cosine similarity of each textual element versus all other textual elements within

^hPlease note that the model does not see any final punctuation as they are omitted in the annotation stage. Hence, the model cannot rely on question marks.

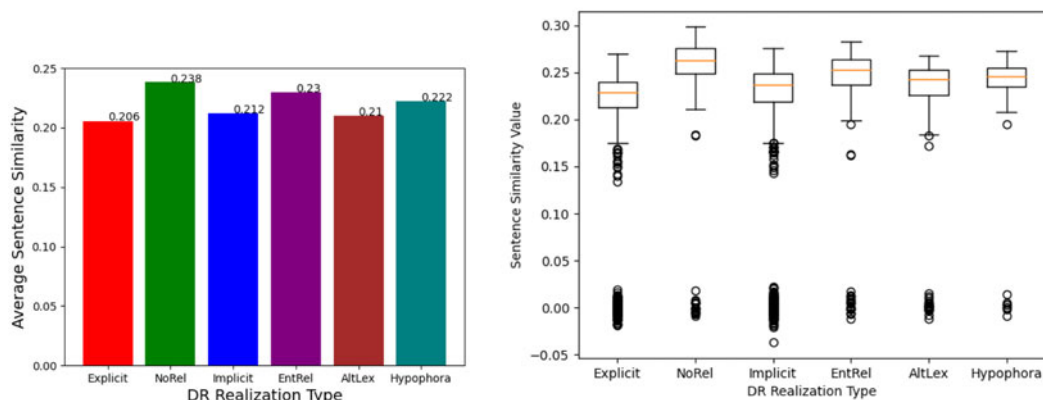


Figure 2. The bar chart and the box plot showing the semantic similarity analysis of DR realization types in the TDB 1.2 (encoded with the Turkish BERT)

each set of similarly realized relation types. Each DR realization type received $(n \times n)$ number of similarity scores, where n is the number of relations in a category.

In Fig. 2, the similarity scores of each DR realization type are plotted in a box plot and their category-wise averages are shown in a bar chart. The box plot shows that the upper and lower quartiles are very close to each other in all categories; most of the scores are squeezed in very small ranges, and even small differences would be significant for our assessment. The analysis shows that the average similarity scores of NoRels, EntRels, and Hypophora are significantly higher than those of other relation realization types. The higher semantic similarity rates mean that they have an identifiable pattern and could be better distinguished. But, EntRels and NoRels have significantly low performance as clearly shown in the confusion matrix (see Fig. C1). The results of this analysis are in line with our earlier observation and confirm that the model cannot identify a clear pattern associated with EntRels and NoRels, which could be due to the low number of annotated EntRel and NoRel tokens.

5.1.2 Sense classification

The sense classification task focuses on the disambiguation of the Level-1 sense conveyed by a discourse relation. Unlike the previous task, sense classification is not performed over all relations. We train separate sense classifiers for explicit and implicit discourse relations, following the common practice. Ideally, one could also train one for AltLex relations as they also convey a sense but that was not possible for the TDB 1.2, due to the lack of enough training data (there is a total of 146 AltLexes in the TDB 1.2 (see Table 1)).

The results of sense classification experiments are provided in Table 5. As expected, sense-wise, explicitly conveyed discourse relations are much easier to classify than implicitly conveyed ones. The explicit sense classifier achieves almost two times better performance than its implicit counterpart (0.41 vs 0.82 Macro-F1) and the classification is very stable across four major sense categories: The classifier achieves 0.81+ F1-Score for each major sense category, which suggests that it is robust to the uneven distribution of labels in the training data (e.g., Expansion relations are twice as frequent as Contingency relations¹). Hence, it should be safe to conclude that such explicit sense classifiers require only several hundreds of examples per label to achieve steady performance.

The sense classification of implicit discourse relations is a notoriously challenging task and often regarded as the most difficult step in shallow discourse parsing. Accordingly, we achieve

¹See Table 2.

Table 5. Level-1 Sense classification results of explicit and implicit DRs in the TDB 1.2

Level-1 Sense	Explicit			Implicit		
	F1-score	Recall	Precision	F1-score	Recall	Precision
Comparison	0.85	0.82	0.89	<u>0.13</u>	0.12	0.17
Contingency	<u>0.81</u>	0.80	0.83	0.37	0.33	0.48
Expansion	<u>0.81</u>	0.76	0.87	0.69	0.72	0.66
Temporal	0.82	0.91	0.74	0.45	0.50	0.43
Accuracy	82.20%			53.37%		
Macro-F1	0.82			0.41		
Weighted F1	0.82			0.54		

much lower performance in the disambiguation of implicit relations. However, considering the size of the training data, the results are in line with our expectations; for example in the PDTB 2.0, the same setup achieves the F1-Score of 0.52 (Kim *et al.* 2020). Unlike explicit relations, the performance varies significantly across different labels, where expansion relations are clearly more successfully classified. This variation can be partly explained by the label distribution in the data: Expansion relations occur almost four times more frequently than the second most frequent label (Contingency). However, the frequency of labels do not explain the poor performance on the Comparison relations since Temporal relations are classified much better despite being slightly less frequent in the training set. Therefore, in addition to the insufficient exposure to some labels, certain relations may be inherently more challenging to classify.

5.1.3 Cross-domain experiments

The lack of annotated data in discourse parsing does not only make it challenging to train high-performance discourse parsers, but, it also hinders the models from generalizing across different domains (Stepanov and Riccardi 2014). Therefore, it is crucial to evaluate the performance of the models on different domains in order to get a complete picture of their real performance. To this end, in this section, we report the performance of our models on the T-TED-MDB. The TED-MDB involves the annotated transcripts of spoken language, and the annotated talks differ from each other in terms of their subject matter. These make this corpus the perfect candidate for such a cross-domain evaluation. The performance of our type and sense classifiers are provided in Tables 6 and 7.

As expected, the performance of all our three models drop on the T-TED-MDB. Yet, the performance drop is within acceptable margins for both explicit (0.82 vs. 0.73 Macro-F1) and implicit (0.41 vs. 0.35 Macro-F1) sense classification. On the other hand, genre change seems to have affected the DR realization type classifier considerably, where the performance drops almost 0.25 in Macro-F1- Score. The performance drops over all DR realization types, but the AltLexes suffer the most significant performance loss, where the classification performance almost drops to half. Considering that AltLexes are rather an open class, compared to explicit connectives, it seems that the models do not learn AltLexes well enough to generalize over unseen AltLex phrases. Overall, although there is much room for improvement, considering the size of the T-TED-MDB, the cross-domain performance of our models is in line with our expectations and can indeed be useful in mining relations on different text types.

5.2 Cross-lingual transfer experiment

The overarching problem in discourse parsing studies in general is the data bottleneck. Although the results provided in the previous sections are on similar levels with what is achieved for English

Table 6. Cross-domain DR realization type classification results over the T-TED-MDB

DR realization type	F1-score	Recall	Precision
AltLex	0.37	0.42	0.35
EntRel	0.18	0.18	0.34
Explicit	0.77	0.80	0.74
Hypophora	0.11	0.20	0.10
Implicit	0.47	0.46	0.49
NoRel	<u>0.09</u>	0.06	0.35
Accuracy	59.85%		
Macro-F1	0.33		
Weighted F1	0.54		

Table 7. Level-1 Sense classification results of explicit and implicit DRs in the T-TED-MDB

Level-1 Sense	Explicit			Implicit		
	F1-Score	Recall	Precision	F1-Score	Recall	Precision
Comparison	0.71	0.69	0.75	0.16	<u>0.14</u>	0.21
Contingency	0.75	0.73	0.77	0.21	0.36	0.17
Expansion	0.80	0.78	0.82	0.65	0.61	0.76
Temporal	<u>0.65</u>	0.83	0.55	0.41	0.40	0.56
Accuracy	73.18%			49.75%		
Macro-F1	0.73			0.35		
Weighted F1	0.76			0.52		

despite the significantly smaller size of the TDB 1.2, clearly, there is much room for improvement. In this second set of experiments, we aimed to explore whether enriching our training data with a larger resource in another language can help us to improve our scores. To this end, we focused on the DR realization type sub-task and performed a cross-lingual transfer experiment using the PDTB 3.0 with the multilingual BERT (mBERT)^j as the text encoder.

In total, we have considered two different settings: (i) the zero-shot setting where mBERT is trained only on the PDTB 3.0 dataset, (ii) the cross-lingual setting, where mBERT is trained on the aggregation of the PDTB 3.0 with the TDB 1.2. We compare the performance of the cross-lingual transfer experiments to the monolingual baseline, where the Turkish BERT model is trained on the TDB 1.2 as discussed in Section 5.1.

The results in Table 8 suggest that the DR realization type classification is a highly language-specific task. In the zero-shot scenario, where the model is exposed to only English examples, the model achieves 72% accuracy and 0.71 weighted F1-Score. Despite still being significantly higher than the chance baseline, the results demonstrate a noticeable decrease when compared to the monolingual baseline. Specifically, the less frequent types experience a significant negative impact, with the performance dropping to zero for Hypophora and NoRel relations, and to half

^j<https://huggingface.co/bert-base-multilingual-cased>

Table 8. F1-Scores of DR realization type classification experiments with Turkish BERT on the TDB 1.2 and with mBERT on both the PDTB 3.0 and the joint dataset

DR realization type	I (TDB 1.2)	II (PDTB 3.0)	III (TDB 1.2 + PDTB 3.0)
AltLex	0.63	0.29 (−0.34)	0.48 (−0.15)
EntRel	0.32	0.47 (+0.15)	0.18 (−0.14)
Explicit	0.90	0.80 (−0.10)	0.87 (−0.03)
Hypophora	0.74	0.00 (−0.74)	0.60 (−0.14)
Implicit	0.77	0.72 (−0.05)	0.71 (−0.06)
NoRel	0.11	0.00 (−0.11)	0.12 (+0.01)
Accuracy (%)	73.90	72.00 (−1.9)	73.64 (−0.26)
Macro-F1	0.58	0.38 (−0.20)	0.49 (−0.09)
Weighted F1	0.77	0.71 (−0.06)	0.72 (−0.05)

(I) Turkish BERT is fine-tuned with the TDB 1.2, and its test set is classified (repetition of Table 4 in Section 5.1.1).

(II) mBERT is fine-tuned with the PDTB 3.0 for zero-shot performance, and the test set of TDB 1.2 is classified.

(III) mBERT is fine-tuned with the joint dataset, PDTB 3.0 + TDB 1.2, and the test set of TDB 1.2 is classified to detect the effect of multilingual training data—cross-lingual transfer learning.

Note: All the differences are calculated from Column I.

for AltLex relations. On the other hand, the performance drop in implicit and explicit relations is not as sharp compared to the other types and can be considered to remain within a reasonable margin. The only performance improvement is detected for EntRel relations.

When the training data compose of both the PDTB 3.0 and the TDB 1.2, the results increase perceptibly as compared to zero-shot setting. The performance becomes close to the monolingual baselines, with only 0.26% loss in accuracy. However, the performance still lags behind the monolingual baselines, meaning that the addition of another language to the training set does not lead to any improvement.

The confusion matrices in Appendix C summarize the classification performance of the monolingual and the multilingual models. A κ (Kappa) statistic is computed to compare the results. We find that the multilingual model has reached a κ coefficient that is 0.025 less than that of our baseline monolingual model, confirming that the performance of the multilingual model lags behind the monolingual model.

Our results mimic the findings of a very recent work on connective prediction, where the authors also report that the language-specific models trained on the target language outperform the multilingual model trained on a concatenation of different languages (Muermans and Kosseim 2022).

Table 8 and our Kappa analysis suggest that the types by which discourse relations are realized in Turkish and English are diverse enough to prevent any knowledge transfer even between the resources that are annotated following the same framework. In order to shed more light into these discrepancies between languages, we have performed a manual error analysis of the predictions of the multilingual model that is trained on both resources. According to our preliminary analysis, the following points stand out as the possible reasons behind the poor generalization across languages:

- Discrepancies between languages in expressing AltLexes: The largest performance drop occurs in AltLex relations, which are open class by definition. Hence, the languages considerably vary when it comes to AltLexes, an observation also put forward by (Özer *et al.* 2022).

- EntRels manifest a performance drop almost as large as AltLexes, and they are frequently confused with implicits and NoRels. These are quite similar to the errors in monolingual experiments (see Section 5.1.1), showing that neither the monolingual model nor the multilingual one learns the EntRel and NoRel labels properly.
- Discrepancies between languages expressing explicit relations: There is a performance drop, albeit small, in the prediction of explicit relations, and our manual error analysis shows that one of the reasons of performance loss is due to the intra-sentential relations conveyed by converbial suffixes, annotated as a type of explicit connectives in Turkish. For example, in sentence 14, despite the presence of the suffixal connective *-se* "if," the relation is mislabeled as an implicit relation.

(14) **Üç kişi versek**, güç çevreleriz.
(**If three of us come together**, we could hardly encircle it.)

English and Turkish annotate discourse connectives belonging to different syntactic classes (e.g., single words vs suffixes), and this could be one reason why the success of cross-lingual transfer decreases in our experiments.

6. Summary and conclusion

Shallow discourse parsing is an important step toward discourse understanding and a prominent contribution to NLU research in general. However, despite its importance, most of the existing work is still confined to English, leaving the field largely understudied in the non-English context. In this paper, we aimed to help remedy this problem by performing various sub-tasks of the shallow discourse parsing pipeline on Turkish. Although our work falls short of developing a full end-to-end parser, it constitutes the most exhaustive study on Turkish so far. Specifically, we focused on a rather overlooked task—the classification of discourse relation realization types, which focuses on understanding whether there is a discourse relation in a given text piece, and if so, how it is realized. Besides, we also performed the well-known tasks of sense classification of explicit and implicit discourse relations.

All tasks are modeled as multi-class text classification problems, and we used a Turkish BERT model to encode the textual elements of a discourse relation. The results suggest that despite the scarcity of the available training data, all tasks can be performed with a satisfactory accuracy, on a par with the reported results on English. In order to gain insight into the DR realization type classification decisions of our monolingual PLM, we first analyzed the results in terms of confusion matrices. Then, using the same PLM, we extracted sentence embeddings by concatenating the arguments of discourse relations and conducted a cosine similarity analysis over the textual elements within DR realization types.

We finally performed a cross-lingual investigation to see if it is possible to leverage information from the much bigger resource of the PDTB 3.0 on the DR realization type classification. In both zero-shot and fully supervised cross-lingual transfer experiments, the performance deteriorated compared to the monolingual baseline, suggesting that even smaller amounts of annotated data in the target language are a better approach than relying on cross-lingual transfer.

Acknowledgements. We would like to thank our anonymous reviewers for insightful suggestions. Any errors that remain are our own.

References

Akkaya E.K. and Can B. (2021). Transfer learning for Turkish named entity recognition on noisy text. *Natural Language Engineering* 27(1), 35–64.

- Alsaif A. and Markert K.** (2011). Modelling discourse relations for Arabic. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP, John McIntyre Conference Centre, A meeting of SIGDAT, a Special Interest Group of the ACL*, 27-31 July 2011, Edinburgh, UK, pp. 736–747.
- Asher N.** (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: Springer Netherlands, Kluwer.
- Bahdanau D., Kyunghyun C. and Yoshua B.** (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Baker C.F., Fillmore C.J. and Lowe J.B.** (1998). The Berkeley FrameNet project. In *Proceedings of 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics, COLING-ACL'98*, August 10-14, 1998, Université de Montréal, Morgan Kaufmann Publishers/ACL, pp. 86–90.
- Başbüyük K. and Zeyrek D.** (2023). Usage disambiguation of Turkish discourse connectives. *Language Resources and Evaluation*, 57(1), 223–256. <https://doi.org/10.1007/s10579-022-09614-3>.
- Bos J.** (2013). The Groningen Meaning Bank. In *Proceedings of the Joint Symposium on Semantic Processing Textual Inference and Structures in Corpora, JSSP 2013*, November 20-22, 2013, Trento, Italy, ACL, p. 2.
- Brown P.F., DeSouza P.V., Mercer R.L., Pietra V.J.D. and Lai J.C.** (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 467–479.
- Byrt T., Bishop J. and Carlin J.B.** (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46(5), 423–429.
- Caselli T. and Üstün A.** (2019). There and back again: Cross-lingual transfer learning for event detection. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, CEUR 2019*, Trento, Italy, CEUR Workshop Proceedings (CEUR-WS.org), vol. 2481, CEUR Workshop Proceedings (CEUR-WS.org).
- CoNLL 2015 Shared Task: Shallow Discourse Parsing** (2015). (accessed September 30, 2020), Available at: <https://www.cs.brandeis.edu/~clp/conll15st/>
- Çakıcı R., Steedman M. and Bozşahin C.** (2018). Wide-coverage parsing, semantics, and morphology. In: Oflazer, K., Saraçlar, M. (eds) *Turkish Natural Language Processing. Theory and Applications of Natural Language Processing*. Springer, Cham. https://doi.org/10.1007/978-3-319-90165-7_8
- Dauphin Y.N., Tür G., Tür D.H. and Heck L.P.** (2014). Zero-shot learning and clustering for semantic utterance classification. In *2nd International Conference on Learning Representations, ICLR 2014*, April 14-16, 2014, Banff, AB, Canada, Conference Track Proceedings, Conference Track Proceedings.
- Devlin J., Chang M.W., Lee K. and Toutanova K.** (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, June 2-7, 2019, Minneapolis, MN, USA, ACL, pp. 4171–4186.
- DISRPT19** (Discourse Unit Segmentation Across Formalisms 2019) (2019). (accessed September 30, 2020). Available at: <https://sites.google.com/view/disrpt2019/shared-task>.
- DISRPT21** (Discourse Unit Segmentation Across Formalisms 2021) (2021). (accessed December 30, 2021). Available at: <https://sites.google.com/georgetown.edu/disrpt2021>.
- Eryiğit G., Nivre J. and Oflazer K.** (2008). Dependency parsing of Turkish. *Computational Linguistics* 34(3), 357–389.
- Fliss J.L.** (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5), 378–382.
- Funaki R. and Nakayama H.** (2015). Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, September 17-21, 2015, Lisbon, Portugal, ACL, pp. 585–590.
- Gopalan S. and Devi S.L.** (2016). BioDCA Identifier: a system for automatic identification of discourse connective and arguments from biomedical text. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining, BioTxtM@COLING 2016*, December 12, 2016, Osaka, Japan, pp. 89–98.
- Grandini M., Bagli E. and Visani G.** (2020). Metrics for Multi-Class Classification: An Overview, ArXiv. /abs/2008.05756.
- Johnson M., Schuster M., Le Q. V., Krikun M., Wu Y., Chen Z., Thorat N., Viégas F., Wattenberg M., Corrado G., Hughes M., Dean J.** (2017). Google's multilingual neural machine translation system: enabling zero-shot translation. *Transactions of the ACL* 5, 339–351.
- Kim N., Feng S., Gunasekara C. and Lastras L.** (2020). Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the ACL*.
- Kishimoto Y., Murawaki Y. and Kurohashi S.** (2020). Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In European Language Resources Association, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*, May 11-16, 2020, Marseille, France, pp. 1152–1158, European Language Resources Association.
- Landis J.R. and Koch G.G.** (1977). The measurement of observer agreement for categorical data. *Wiley, International Biometric Society* 33(1), 159–174, 1977.
- Levy O., Seo M., Choi E. and Zettlemoyer L.** (2017). Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning, (CoNLL 2017)*, August 3-4, 2017, Vancouver, Canada. ACL, pp. 333–342.
- Liang L., Zhao Z. and Webber B.** (2020). Extending implicit discourse relation recognition to the PDTB 3.0. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, November 2020. ACL, pp. 135–147.

- Lin Z., Kan M.Y. and Ng H.T.** (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, A Meeting of SIGDAT, a Special Interest Group of the ACL*, 6-7 August 2009, Singapore, pp. 343–351.
- Lin Z., Ng H.T. and Kan M.Y.** (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering* 20(2), 151–184.
- Loshchilov I. and Hutter F.** (2017). Decoupled weight decay regularization. ArXiv./abs/1711.05101.
- Luong T., Pham H. and Manning C.D.** (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. ACL, pp. 1412–1421.
- Ma Y., Cambria E. and Gao S.** (2016). Label embedding for zero-shot fine-grained named entity typing. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, December 11-16, 2016, Osaka, Japan. ACL, pp. 171–180.
- Marcu D. and Echihiabi A.** (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of ACL*, July 6-12, 2002, Philadelphia, PA, USA. ACL, pp. 368–375.
- Muermans T.C. and Kosseim L.** (2022). A BERT-based approach for multilingual discourse connective detection. In *Proceedings, Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022*, vol. 13286. Valencia, Spain: Springer Nature, pp. 449.
- Mukherjee S., Tiwari A., Gupta M. and Singh A.K.** (2015). Shallow discourse parsing with syntactic and (a few) semantic features. In *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015*, July 30-31, 2015, Beijing, China. ACL, pp. 61–65.
- Muller P., Braud C. and Morey M.** (2019). ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, June 2019, Minneapolis, MN. ACL, pp. 115–124.
- Nie A., Bennett E. and Goodman N.D.** (2019). Dissent: Learning sentence representations from explicit discourse relations. In Long Papers, *Proceedings of the 57th Conference of ACL, ACL 2019*, July 28-August 2, 2019, Florence, Italy. ACL, vol. 1, pp. 4497–4510, Long Papers.
- Oflazer K. and Bozşahin H. C.** (1994). Turkish natural language processing initiative: An overview. In *Proceedings of the Third Turkish Symposium on Artificial Intelligence and Artificial Neural Networks*. Ankara.
- Oflazer K. and Saraçlar M.** (2018). Turkish and its challenges for language and speech processing. In *Turkish Natural Language Processing*. Springer, pp. 1–19.
- Özer S., Kurfalı M., Zeyrek D., Mendes A. and Valūnaitė Oleškevičienė G.** (2022). Linking discourse-level information and the induction of bilingual discourse connective lexicons. *Semantic Web* 13(6), 1081–1102. <https://doi.org/10.3233/SW-223011>.
- Palmer M., Gildea D. and Kingsbury P.** (2005). The Proposition Bank: an annotated corpus of semantic roles. *ACL* 31(1), pp. 71–106.
- Pasupat P. and Liang P.** (2014). Zero-shot entity extraction from web pages. In Long Papers, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, June 22-27, 2014, Baltimore, MD, USA. ACL, vol. 1, pp. 391–401, Long Papers.
- Pitler E., Louis A. and Nenkova A.** (2009). Automatic sense prediction for implicit discourse relations in text. In *ACL 2009, Proceedings of the 47th Annual Meeting of ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2-7 August 2009, Singapore, pp. 683–691.
- Pitler E. and Nenkova A.** (2009). Using syntax to disambiguate explicit discourse connectives in text. In Short Papers, *ACL 2009, Proceedings of the 47th Annual Meeting of ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2-7 August 2009, Singapore, pp. 13–16.
- Pitler E., Raghupathy M., Mehta H., Nenkova A., Lee A. and Joshi A.** (2008). Easily identifiable discourse relations. In *Coling 2008: Companion Volume: Posters*. Manchester, UK. Coling 2008 Organizing Committee, pp. 87–90.
- Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A. and Webber B.** (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.
- Prasad R., Miltsakaki E., Dinesh N., Lee A. and Joshi A.** (2008). The Penn Discourse TreeBank 2.0 Annotation Manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.
- Prasad R., Webber B. and Joshi A.** (2014). Reflections on the Penn Discourse Treebank: comparable corpora, and complementary annotation. *ACL* 40(4), pp. 921–950, 2014.
- Prasad R., Webber B. and Lee A.** (2018). Discourse annotation in the PDTB: the next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, August 2018, Santa Fe, New Mexico, USA. ACL, pp. 87–97.
- Ramesh B.P., Prasad R., Miller T., Harrington B. and Yu H.** (2012). Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association* 19(5), 800–808.
- Rutherford A. and Xue N.** (2014). Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of ACL, EACL 2014*, April 26-30, 2014, Gothenburg, Sweden, pp. 645–654.

- Schuster S., Gupta S., Shah R. and Lewis M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. In (Long and Short Papers), *Proceedings of the 2019 Conference of the North American Chapter of ACL: Human Language Technologies, NAACL-HLT 2019*, June 2-7, 2019, Minneapolis, MN, USA, 1, pp. 3795–3805, (Long and Short Papers).
- Seker G.A. and Eryiğit G. (2017). Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content. *Semantic Web* 8(5), 625–642.
- Spooren W. and Degand L. (2010). Coding coherence relations: reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2), 241–266.
- Stepanov E. and Riccardi G. (2014). Towards cross-domain PDTB-style discourse parsing. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*.
- Tallón-Ballesteros A.J. and Riquelme J.C. (2014). Data mining methods applied to a digital forensics task for supervised machine learning. In *Studies in Computational Intelligence*, pp. 413–428, 2014.
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L. and Polosukhin I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, December 2017, Long Beach, CA, USA, pp. 5998–6008.
- Wang W., Zheng V.W., Yu H. and Miao C. (2019). A survey of zero-shot learning: settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology* 10(2), 37–37.
- Warrens M.J. (2014). New Interpretations of Cohen's Kappa, *Journal of Mathematics*, Hindawi Publishing Corporation. Available at <https://www.hindawi.com/journals/jmath/2014/203907/>.
- Webber B., Prasad R., Alan L. and Joshi A. (2019). The Penn Discourse Treebank 3.0 Annotation Manual. Technical report. Institute for Research in Cognitive Science, University of Pennsylvania.
- Webber B., Prasad R. and Lee A. (2019). Ambiguity in explicit discourse connectives. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, May 2019, Gothenburg, Sweden. ACL, pp. 134–141.
- Webber B., Prasad R., Lee A. and Joshi A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, August 2016, Berlin, Germany. ACL, pp. 22–31.
- Xue N., Ng H.T., Pradhan S., Prasad R., Bryant C. and Rutherford A.T. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015*, July 30-31, 2015, Beijing, China. ACL, pp.1–16.
- Yazdani M. and Henderson J. (2015). A model of zero-shot learning of spoken language understanding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, September 17-21, 2015, Lisbon, Portugal. ACL, pp. 244–249.
- Zeldes A., Das D., Maziero E.G., Antonio J.D. and Irukieta M. (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*, June 2019, Minneapolis, MN. ACL, pp. 97–104.
- Zeyrek D. and Başbüyük K. (2019). TCL - A Lexicon of Turkish discourse connectives. In *Proceedings of the First International Workshop on Designing Meaning Representations*, August, 2019, Florence, Italy. ACL, pp. 73–81.
- Zeyrek D. and Er M.E. (2022). A description of Turkish Discourse Bank 1.2 and an examination of common dependencies in Turkish Discourse. In *Proceedings of The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing, ALTNLP'22*, June 7-8, 2022, Koper, Slovenia, ceur-ws.org/Vol-3315/paper04.pdf.
- Zeyrek D. and Kurfalı M. (2017). TDB 1.1: Extensions on Turkish Discourse Bank. In *Proceedings of the 11th Linguistic Annotation Workshop, LAW@EACL 2017*, April 3, 2017, Valencia, Spain. ACL, pp. 76–81.
- Zeyrek D. and Kurfalı M. (2018). An assessment of explicit inter- and intra-sentential discourse connectives in Turkish Discourse Bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, May 7-12, 2018, Miyazaki, Japan. European Language Resources Association (ELRA).
- Zeyrek D., Mendes A., Grishina Y., Kurfalı M., Gibbon S. and Ogrodniczuk M. (2020). TED multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation* 54(2), 587–613.
- Zeyrek D., Mendes A. and Kurfalı M. (2018). Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Zeyrek D. and Webber B. (2008). A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources*.
- Zhao Z. and Webber B. (2021). Revisiting shallow discourse parsing in the PDTB-3: handling intra-sentential implicits. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, November 2021, pp. 107–121. Available at <https://doi.org/10.18653/v1/2021.codi-main.10>.

A. PDTB 3.0 sense hierarchy

Level-1	Level-2	Level-3
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE SUCCESSION
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND
ARG2-AS-COND		
CONDITION+SPEECHACT	–	
NEGATIVE-CONDITION	ARG1-AS-NEGCOND	
	ARG2-AS-NEGCOND	
NEGATIVE-CONDITION+SPEECHACT	–	
PURPOSE	ARG1-AS-GOAL	
	ARG2-AS-GOAL	
COMPARISON	CONCESSION	ARG1-AS-DENIER
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
SIMILARITY	–	
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL
ARG2-AS-DETAIL		
MANNER	ARG1-AS-MANNER	
	ARG2-AS-MANNER	
SUBSTITUTION	ARG1-AS-SUBST	
	ARG2-AS-SUBST	

Figure A1. The leftmost column contains the Level-1 senses and the middle column contains the Level-2 senses. For asymmetric relations, Level-3 senses are located in the rightmost column (Webber *et al.* 2019). While the TDB 1.2 and the PDTB 3.0 both assign senses from all three levels, the present work exploits Level-1 senses only.

B. The classification model, method of fine-tuning and tests

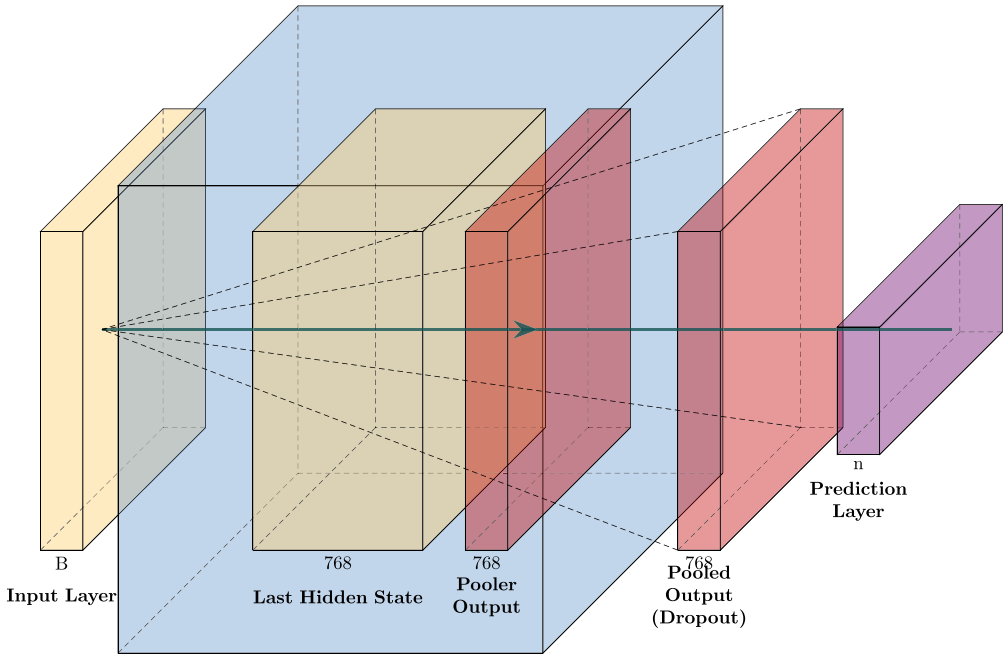
As shown in Fig. B1, a neural network-based classification model is devised and magnified up to 768 hidden layers with 184,345,344 parameters for the improvement of results by the virtue of the GPU memory size, which is fully utilized. The first component of the model is the Input layer that feeds the model with B (Batch Size) number of DRs (encoded by BERT) in each iteration. The main body of the model is TensorFlow BERT, and the Main Layer is *TFBaseModelOutputWithPoolingAndCrossAttentions* class, released in transformers library.^k It forms a base class for the model's outputs, and it also contains a pooling of the last hidden states. Its novelty is the cross-attention mechanism instead of self attention: here, after the attention softmax, the attention weights of the decoder's cross-attention layer are used to compute the weighted average in the cross-attention heads.

The class has two components: The Last Hidden State and the Pooler Output. The Last Hidden State of the TensorFlow BERT Main Layer is a TensorFlow tensor of the shape B, M (maximum sequence length (128)), and H (number of hidden layers (768)). It forms the sequence of hidden states at the output of the last layer of the model. The BERT Main Layer is followed by the Pooler Output, which is a TensorFlow tensor of the shape B and H. The Pooler Output is the last layer where the hidden state of the first token of the sequence (classification token) is further processed by a linear layer and a Tanh activation function. The Pooled Output is the drop out layer of shape B and H, which collects the results for the softmax function. The final layer is the prediction layer of shape n (number of classes) and B, which creates a discrete prediction for each DR by an argmax function.

In all the experiments, the fine-tuning of the specific PLM involves the use of a supervised learning approach, which is literally a training that results in a new PLM to be used for the encoding of the test set. The other aspects of the experiment setup are as follows:

- The classification task is conducted separately for all DR realization types and their Level-1 senses.
- The input is pairs of arguments (Arg1 and Arg2), and the output is a label, such as a DR realization type label or a sense label annotated in the data.
- The input dataset is formed by concatenating the text spans of arguments (Arg1, Arg2) to form the "text" feature for BERT encoding.
- The DR realization type labels and the sense labels form the "category" feature of the input for both training and test phases.
- Hyper-parameter tuning is done empirically by repeating the steps below:
 1. Concatenate the binary arguments of a DR together with the discourse connective (if available) into single lines.
 2. Take 128 as the maximum input sequence length (the maximum number of words that represent each DR). Tokenize that many number of words from the texts with the BERT tokenizer and convert all into the indexes of the tokenizer vocabulary.
 3. Pad or truncate the texts into the maximum length long vectors.
 4. Create an attention mask and return a dictionary of outputs.
 5. Fine-tune the PLM by training it with the *Bert_MultiClass* classification model, depicted in Fig. B1, by using AdamW optimizer (Loshchilov and Hutter 2017) with the learning rate of $5e - 5$.

^khttps://github.com/huggingface/transformers/blob/main/src/transformers/modeling_tf_outputs.py



TFBertMainLayer (TFBaseModelOutputWithPoolingAndCrossAttentions)

Figure B1. The symbolic representation of the BERT MultiClass TensorFlow Model

6. The [CLS] token is used to represent the whole relation. It is classified by a dense layer to reach the final prediction. As for the loss function, we used the cross entropy following the common practice.
- The DRs in the test set are encoded with the fine-tuned model and classified by using the same *Bert_MultiClass* classification model.

C. Kappa analysis over confusion matrices

Our monolingual and multilingual classifier results are presented in the form of confusion matrices in Figs. C1 and C2 below. They provide valuable quantitative values, including the total number of samples, precision, recall, and F1-Score values, but we also need an objective method to compare the performance of our models. For this purpose, we use the Cohen’s Kappa Association Coefficient (κ) (Landis and Koch 1977).

For more than five decades, κ has been used as an associating measure for providing an agreement score between two observers on a nominal scale and its formula is built upon an N by k observation matrix in which the elements n_{ij} represent the number of observers who assigned the i -th case in the j -th class:

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}, \quad P_i = \frac{1}{n(n-1)} \left(\sum_{j=1}^k n_{ij}^2 - n \right) \tag{C1}$$

$$P_o = \frac{1}{N} \sum_{i=1}^N P_i, \quad P_e = \sum_{j=1}^k p_j^2, \tag{C2}$$

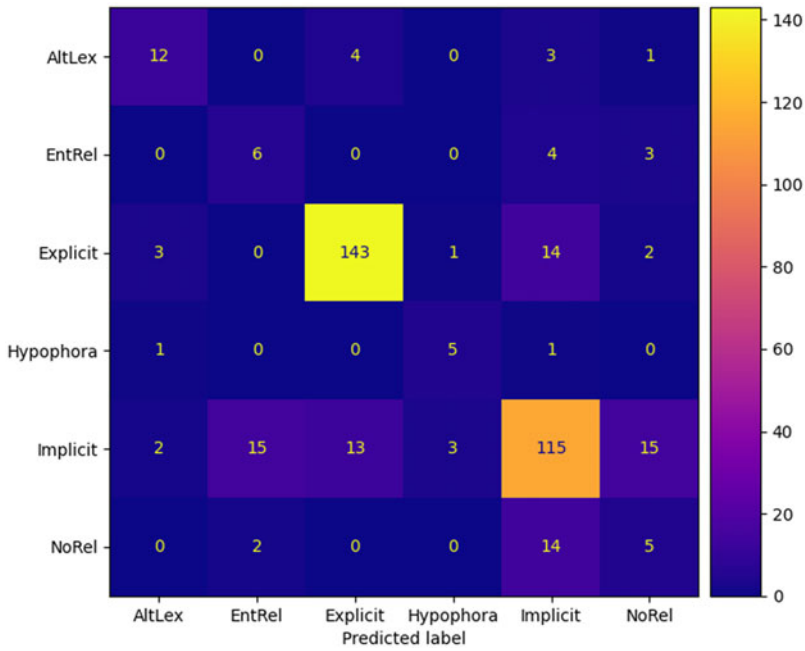


Figure C1. The confusion matrix of DR realization type classification with the TDB 1.2 test set, where the model is trained over TDB 1.2, and encoded with the fine-tuned monolingual BERT

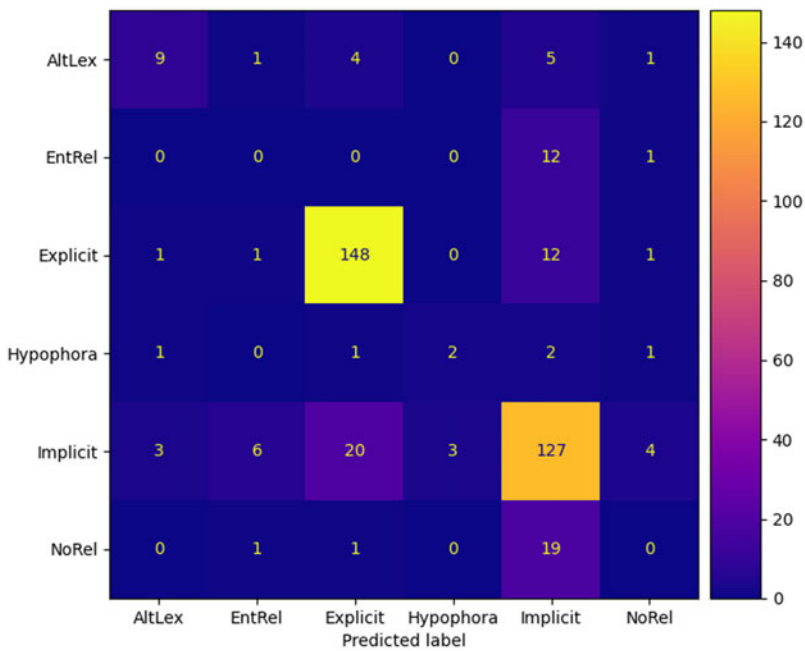


Figure C2. The confusion matrix of DR realization type classification with the TDB 1.2 test set, where the model is trained on the custom multilingual dataset (TDB 1.2 + PDTB 3.0) and encoded with the fine-tuned mBERT

Table C1. κ Coefficient of the confusion matrices in Figs. C1 and C2 calculated by Formula 4

	Monolingual model (TDB 1.2)	Multilingual model (TDB 1.2 + PDTB 3.0)
κ	0.6	0.575
κ Difference		-0.025

where N is the number of all data samples annotated, p_j is the proportion of all assignments to the j -th class, P_i is the extent of agreement among the n observers for the i -th sample, P_o is the observed overall agreement, and P_e is the expected mean proportion of agreement due to chance (Fleiss 1971).

So, the Kappa statistic is defined as the degree of actually attained agreement in excess of chance ($P_o - P_e$), normalized by the maximum agreement attainable above chance ($1 - P_e$) (Grandini, Bagli, and Visani 2020):

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{C3}$$

The κ statistic reduces the ratings of two observers to a single number (Warrens 2014) by taking into account a priori distribution that does not affect the distribution of the predictions among the target classes. The bias between observers and the distribution of data across the categories (prevalence) affect κ in very complex ways (Byrt, Bishop, and Carlin 1993) and κ 's ability to compensate for random hits makes it an interesting alternative for measuring the success levels of classifiers. Especially in working with unbalanced data such as ours, the κ coefficient can be helpful in comparing the performance of classification models. For inter-annotator agreement evaluation, the range of κ coefficients extend from -1 to $+1$ such that -1 , 0 and $+1$ indicate strong disagreement, chance-level agreement and strong agreement, respectively. In the evaluation of classification over confusion matrices, κ coefficients could be interpreted from very poor to perfect classification.

We calculated the κ coefficients of our monolingual and multilingual classification models using the values in the confusion matrices as follows (Tallón-Ballesteros and Riquelme 2014):

$$\kappa = \frac{\sum_{i=1}^m CM_{ii} - \sum_{i=1}^m C_{i_{corr}} \cdot C_{i_{pred}}}{N^2 - \sum_{i=1}^m C_{i_{corr}} \cdot C_{i_{pred}}} \tag{C4}$$

where

- CM_{ii} represents the diagonal elements of the confusion matrix,
- $C_{i_{corr}}$ is the number of correct samples in the i -th class,
- $C_{i_{pred}}$ is the number of predicted samples picked for the i -th class.

We measured the κ coefficients of the values in confusion matrices in Figs. C1 and C2 by the Formula 4. That is, we calculated a random accuracy by taking the sum of all multiplications of the number of correct samples in the i -th class ($C_{i_{corr}}$) with the number of each predicted sample picked for the i -th class ($C_{i_{pred}}$). Then, in order to calculate a κ coefficient for each confusion matrix, we divided the difference between the sum of all true positives (the diagonal elements of the confusion matrix) and random accuracy into the difference between the square of the data sample size in the test (N) and Random Accuracy.

The results are given in Table C1, showing a decrease of 0.025 in the κ coefficient of the multilingual model.