Annotating and Disambiguating the Discourse Usage of the Enclitic *dA* in Turkish

Ebru Ersöyleyen and **Deniz Zeyrek** and **Fırat Öter**
Graduate School of Informatics
Middle East Technical University
Ankara, Türkiye
{ebru.ersoyleyen,dezeyrek,foter}@metu.edu.tr

Abstract

The Turkish particle *dA* is a focus-associated enclitic, and it can act as a discourse connective conveying multiple senses, like additive, contrastive, causal etc. Like many other linguistic expressions, it is subject to usage ambiguity and creates a challenge in natural language automatization tasks. For the first time, we annotate the discourse and non-discourse connnective occurrences of *dA* in Turkish with the PDTB principles. Using a minimal set of linguistic features, we develop binary classifiers to distinguish its discourse connective usage from its other usages. We show that despite its ability to cliticize to any syntactic type, variable position in the sentence and having a wide argument span, its discourse/non-discourse connective usage can be annotated reliably and its discourse usage can be disambiguated by exploiting local cues.

1 Introduction

Discourse connectives are one of the most important aspects of discourse structure. They are lexico-syntactic elements that signal a pragmatic or semantic relation (contingency, expansion, contrast, etc.) between two discourse units such as verb phrases, clauses or sentences (Asher, 1993; Prasad et al., 2008). While the most well-known discourse connectives belong to syntactic classes such as coordinating and subordinating conjunctions (*and*, *but*, *because*), adverbs (*however*) or prepositional phrases (*in sum*), it is known that clitics can also function similarly as discourse connectives (König, 2002), and may convey additive, contrastive or concessive senses (Forker, 2016; Faller, 2020).

Clitics are particles that are phonologically dependent on the lexical item to which they are attached and in many languages, they play a role in expressing focus. Usually, all types of phrases (noun phrases, verb phrases, etc.) can function as foci of a particle. In Turkish, too, most clitics are attached to phrases. The enclitic *dA* is a special particle, which is both focus- and topic-associated. In this respect, *dA* (orthographically *"de", "da"*) is even more worth investigating.[1]

The focus-sensitive characteristics and the discourse connective role of *dA* have long been noticed in the Turkish linguistics literature (Kerslake, 1992; Ergin, 1975; Erdal, 2000; Göksel and Özsoy, 2003). However, its discourse connective usage has not been annotated in the existing discourse-level Turkish corpora and, to the best of our knowledge, it has not been the topic of a computational discourse analysis so far. The existing experiments are limited to the disambiguation of the orthographic forms *da* (one of the representations of *dA*) and *-da* (one of the representations of the locative suffix, *-DA*) (Arıkan et al., 2019).[2]

It is known that connectives are susceptible to usage ambiguity, that is "whether or not a given token is serving as a discourse connective in its context" (Webber et al., 2019a), and this has initiated usage disambiguation tasks over connectives in many languages. Well-known works that disambiguate English connectives involve Pitler and Nenkova (2009) and Lin et al. (2010). Similar tasks have been carried out in Chinese (Shih and Chen, 2016), French (Laali and Kosseim, 2016), German (Dipper and Stede, 2006; Schneider and Stede, 2012) and Turkish (Başıbüyük and Zeyrek, 2023). To facilitate Natural Language Processing (NLP) tasks such as text summarization, automatic translation, knowledge extraction, etc., usage ambiguity tasks have to involve clitics as well as other types of discourse connectives. Given that there

[1] The upper case letter "A" is used to represent the alteration of vowels ("-e", "-a") with respect to the last syllable of the preceding word.

[2] *da*, one of the representations of the clitic *dA*, can be misspelled and written as a suffix, in which case it becomes a homograph of one of the variants of the locative suffix *-DA*, i.e. *-da*. This motivates the work on the disambiguation of the orthographical forms. In the current work, the upper case letter "D" represents the alternation of the alveolar consonants ("t" and "d") with respect to consonant assimilation rules.

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 46–54

is a research gap in the usage disambiguation of *dA* in particular and Turkish clitics in general, this paper describes an annotation study followed by a classification task over a corpus where *dA*'s various non-discourse connective roles are distinguished from its discourse connective roles. The summary of our contributions are:

- We construct a reliably annotated dataset of the discourse and non-discourse connective usages of *dA* following the principles of the PDTB (Prasad et al., 2008; Webber et al., 2019b) in terms of discourse connective spotting.
- By using a set of simple linguistic features, we run machine learning experiments to disambiguate the cases where *dA* is used as a discourse connective.
- We show that our basic features can distinguish between *dA* classes to a significant extent.

The outline of the paper is as follows: Section 2 focuses the linguistic behavior of *dA*. It provides a description of its various functions demonstrating its usage ambiguity. In Section 3, the data creation stage of our work is described, the annotation style and inter-annotator agreement results are presented. In Section 4, we describe our experimental setup by introducing the feature set, data processing (e.g. lemmatization, tagging) and the classification algorithms we used. An evaluation of the success of the models and an error analysis are presented. In Section 5, we summarize our work also discussing its limitations and contributions, and offer some ideas for future work.

2 Background

Turkish is a verb-final, agglutinating language, where suffixation plays an important role both in derivation and inflection. It has clitics such as *mI* (the marker of yes/no questions), *(y)DI* and *(y)mIş* (copular markers), and *dA*. An important grammatical fact that teases apart clitics and affixes is that while clitics can attach to material already containing clitics, affixes cannot (Erdal, 2000).

2.1 Basic facts regarding *dA*

As shown by previous researchers, many clitics are multifunctional, and the Turkish *dA* is no different from the clitics in other languages in its multifaceted behaviour. It is basically an additive (akin to English *too, also*, as discussed in the extensive typological study by Forker (2016)). This function goes together with *dA*'s focalizer or intensifier role. Moreover, *dA* has a variable position in the sentence as it can cliticize to any syntactic type and, as it is the case with most of the connectives, it is not easy to demarcate the boundaries of its arguments, i.e. ARG1, ARG2. In the current work, the arguments to *dA* were not annotated.[3]

Syntactically, *dA* is placed at the right outermost boundary of a word to the right of all other case suffixes (Göksel and Özsoy, 2003). Like other clitics mentioned by Zwicky (1977), it cannot be moved independently of its host without change in meaning, but it can be moved with its host as long as the constraints on word order configurations permit. Göksel and Özsoy (2003) show that *dA* can occur with focused or unfocused constituents, but always in a sentence that contains focus. In sentences with *dA*, a set of alternatives is evoked by focus (Rooth, 1992), or by *dA* itself, and *dA* asserts the truth of one of these alternatives. In our work, as also stated by Göksel and Özsoy (2003), we consider *dA* not an additive marker itself; rather, when the presupposition it carries is interpreted together with the rest of the utterance, the additive sense arises.

Throughout the paper, the use of "a" and "b" in the examples denotes the discourse segments linked by a discourse connective. The discourse and non-discourse connective role of the clitic is abbreviated as DC and NDC, respectively. Morpheme-by-morpheme[4] analyses are provided to indicate the variable position of *dA* in the sentence as well as the word to which it clitizes (shown in bold fonts).

2.2 The discourse connective and non-discourse connective usages of *dA*

Following the principles of the PDTB, we consider *dA* a DC when it links two segments that have an "abstract object" interpretation (propositions, eventualities, etc.) (Asher, 1993; Prasad et al., 2008).

[3] In the PDTB framework, the text spans with an abstract object interpretation are legal arguments to a connective; connectives link two text spans with an abstract object interpretation referred to as ARG1 and ARG2.

[4] The morpheme abbreviations we use throughout this article are as follows: ABIL abilitative marker, ACC accusative case, AUX auxiliary, CAUS causative, COND conditional, CV converb marker, DAT dative case, GEN genitive case, MOD modifier, NEG negative, OPT optative, PL plural, POSS possessive, PRES present, PROG progressive, PST past, SG singular, VN verbal noun marker, 1 first person, 2 second person, 3 third person.

In its DC role, it always invites a continuative inference: It allows the extension of information expressed in the first segment by providing further detail in the second segment. The second segment is a "separate but parallel" piece of information (König, 2002) and it is often the host clause for *dA*, as shown in (1) - (2):

(1) a. Sen-i sev-iyor-um de-di,
you-ACC love-PROG-1.SG say-PST
b. **ben** de inan-dı-m.
I *dA* believe-PST-1.SG
'He said 'I love you', and I believed him.' (DC)

(2) a. Halil'in gel-diğ-in-i
Halil-GEN come-VN-2.SG.POSS-ACC
fark et-me-di-ler.
notice do-NEG-PST-3.PL
b. **Halil** de kadınlar-a bir şaka
Halil *dA* ladies-DAT a joke
yap-ma-ya karar ver-di.
do-VN-DAT decide-PST-3.SG
'They did not notice that Halil came, and Halil decided to play a joke on the ladies.' (DC)

dA can also occur in the first segment, attached to the predicate as in examples (3), (4) or an auxiliary, as shown in (5).

(3) a. **Bekle** de,
Wait *dA*,
b. gel-ince konuş.
come-CV speak.
'Wait, and then speak when he comes.' (DC)

(4) a. İyi **güzel** de,
okay nice *dA*
b. bir bak-alım.
have.a look-OPT
'Okay, it's nice, but let's have a look.' (DC)

(5) a. Beni ara-dın **mı** da,
me call-PST.2SG AUX *dA*,
b. yanıt bekli-yor-sun
answer expect-PROG.2SG
'Is it the case that you've called me so you're expecting an answer?' (DC)

In (6), *dA* has a different function, namely, it introduces a new topic rather than conveying a discourse relation. In this excerpt, two friends (A and B) are in an exhibition. Pointing to one of the paintings, A starts the conversation (segment a) and B responds (segment b). We consider *dA* an NDC in this role.

(6) a. Bu **tablo-yu** da Ali al-dı.
This painting-ACC *dA* Ali buy-PST
b. Güzel.
'Nice'.
'A: As for this painting, Ali bought it'.
'B: It's nice'. (NDC)

dA may appear in a discontinous form, acting as a coordinator. In this usage, it often corresponds to the conjunction *both ... and* in English. We consider it a DC in its VP coordination role as illustrated in (7), an NDC otherwise, e.g. when adjectives are coordinated, as in (8).

(7) a. Çocuk kedici-ği **okşa-dı** da
child kitty-ACC caress-PST *dA*
b. **öp-tü** de.
kiss-PST *dA*
'The child both caressed and kissed the kitty.' (DC)

(8) Kız- ın saçlar-ı **kızıl** da **kıvırcık** da.
Girl-GEN hair-POSS red *dA* wavy *dA*
'The girl's hair is both **red** and **frizzly**.' (NDC)

In addition to these, *dA* can cliticize to conjunctions and adverbs, yielding the emphatic form of that conjunction or adverb. For example, *ve de* 'and *dA*' is the emphatic form of *ve* 'and' (9). In these cases, the head of the discourse relation (*ve*) is considered the discourse connective and *dA* its modifier (Zeyrek et al., 2013). That is, we do not consider *dA* as the sole discourse connective in such cases and mark it as NDC.

(9) a. Komik ol-malı-yım,
Funny be-ABIL-PRES-1.SG
gül-dür-meli-yim
laugh-CAUS-ABIL-PRES-1.SG
b. **ve** de aşık
and MOD love
ol-malı-yım.
fall.in.love-ABIL-PRES-1.SG
'I should be funny, make [people] laugh; and furthermore, I should fall in love.' (NDC)

On the other hand, *dA* also cliticizes to the conditional, (*y*)*sA* (10), contributing a concessive sense to the sentence, akin to the role of 'even though'. Thus, we annotate its use with (*y*)*sA* as DC in examples like (10).

(10) a. Aynı öneri-yi
same suggestion-ACC
sun-du-k-**sa** da,
offer-PST-1.PL-COND *dA*

b. yanaş-ma-dı-lar.
accept-NEG-PST-3.PL
'Even though we made the same suggestion, they didn't accept it.' (DC)

Forbes-Riley et al. (2006) show that there are clausal adverbs (*probably, usually*) and discourse adverbials (*as a result, in addition, consequently*). These are semantically different forms; while clausal adverbs are interpretable with respect to just their matrix clause, discourse adverbials require an abstract object interpretation from prior discourse. So, clausal adverbs are not discourse connectives. *dA* can cliticize to clausal adverbs such as *belki de* 'perhaps *dA*', *gerçekten de* 'indeed *dA*' (11). It can also attach to discourse adverbs (*özellikle de* 'in particular *dA*'), but it is always considered a modifier (hence NDC) when it is cliticized to an adverb.

(11) **Gerçekten** de onun eli açık-tı.
Indeed MOD his hand open-PST
'Indeed *dA*, he was very generous.' (NDC)

3 Data Construction and Reliability Analysis

3.1 Data

To build a corpus for the current study, we started with the TDB 1.1 (Zeyrek and Kurfalı, 2017), an annotated corpus of explicit and non-explicit discourse connectives, their binary arguments and senses in the PDTB 2.0 style (Prasad et al., 2008). Due to having linguistic characteristics quite different than other connectives such as conjunctions and adverbs, *dA* was not systematically annotated in the TDB 1.1; its analysis was postponed until a new annotation study that solely focuses on clitics or *dA* itself could be launched. For the current work, we had initially planned to work on the TDB to extend it with a systematic annotation of *dA*.

A manual inspection of the TDB 1.1 showed that it does not have an adequate number of discourse and non-discourse *dA* occurrences. We decided to create a new dataset, referred to as the *dA Corpus*, by combining selected *dA* samples from the TDB 1.1 with those extracted from another Turkish corpus, namely, the TS Corpus v2 (Sezer and Sezer, 2013; Sezer, 2017).[5] Table 1 shows the distribution of the sources of selected samples in the corpus.[6]

# samples with *dA*	
TDB 1.1	TS Corpus v2
436	438

Table 1: The *dA* corpus.

3.2 Annotation Style and Inter-annotator Agreement (IAA)

Discourse connectives are clear signals that show how discourse units are linked by a pragmatic/semantic relationship. Taking this description and the PDTB 2.0 annotation guidelines as our starting points, we wrote a set of guidelines describing how to recognize the discourse connective and non-discourse connective uses of *dA* mentioned in the current paper. Each sample minimally contained clauses to the immediate right and left context of *dA*, but there were samples that had more than one clause on each side, as they were deemed necessary to infer the meaning of the text. Since a piece of text may have multiple *dA* instances, the tokens to be annotated were highlighted. All the *dA* samples were annotated by two independent, native speaker annotators. They were asked to annotate all the text pieces where the clitic is highlighted.

Although the annotation of *dA*'s discourse senses is out of our scope, we asked our annotators to pay attention to the senses of *dA* when they infer a discourse relation made salient by *dA*. The annotators were told that the basic sense of *dA* is to indicate addition.[7] They were also told that they may infer additional senses such as temporal succession (3), concession (4), result (5).[8] Since *dA* lacks such

[5]The TDB 1.1 is a 40.000-word, multi-genre (research surveys, articles, interviews, news articles, novels), written corpus of modern Turkish. The TS Corpus is based on the BOUN Web Corpus (Sak et al., 2008), containing data from news and other internet websites. It is composed of over 491M units, where all units are marked on the basis of word type (POS tag), morphological structure tag (Morphological Tagging) and root word (Lemma).

[6]The corpus is available at https://github.com/TurkishdA/dA-Corpus.

[7]In the current work, the additive discourse sense corresponds to Expansion.Conjunction or Expansion.Detail.Arg2-as-detail senses in the PDTB 3.0.

[8]In cases like (3) a sense in addition to the additive sense is inferred. These are a type of multiple relations introduced in the PDTB 3.0. In other examples such as (4) and (5), *dA* conveys a single sense. But the annotators are not required to differentiate between single sense versus multiple senses of *dA* tokens, which is left for further work.

additional senses in its non-discourse connective roles, to notice them in the data would further help the annotators while tagging its DC usage.

The annotation cycle involved two steps. First, annotator1 annotated the entire occurrences of *dA* as DC or NDC. In two sessions, which lasted approximately two hours, the guidelines were explained to the independent annotator and a few examples that are not involved in the data were annotated jointly; then annotator2 tagged all the *dA* occurrences highlighted in the corpus.

To measure the inter-annotator agreement, we adopted the method used in Zeyrek et al. (2020) and took one set of annotations (namely those created by annotator1) as the correct annotations since the annotations were created by one of the members of our research team. We calculated the IAA with the standard metrics of Precision, Recall and F1 in formulas[9] 1, 2 and 3, respectively. The results are presented in Table 2.

We also evaluated the IAA with the kappa statistic (Cohen, 1960) to assess base level agreement. The result showed a substantial agreement between annotators with a κ score of 0.74.

$$Precision = \frac{\#\ of\ correct\ DC\ assg.s}{\#\ of\ DC\ assg.s} \quad (1)$$

$$Recall = \frac{\#\ of\ correct\ DC\ assg.s}{\#\ of\ DC\ samples} \quad (2)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

Precision	Recall	F-score
0.89	0.86	0.87

Table 2: IAA results.

Once we obtained the inter-annotator results, in the second step, we spotted and discussed the disagreed cases in a series of meetings, and reached a unanimous agreement as to whether a disagreed *dA* token is DC or NDC. We thus obtained the gold standard data. Table 3 provides the number of adjudicated DC and NDC tokens in the *dA* corpus. Table 4 lists the different word classes to which *dA* cliticizes and their frequencies (see Appendix A for the distribution of POS tags across DC/NDC instances).

[9] *Assignments* is abbreviated as *assg.s.*

# DC	# NDC
473	401

Table 3: Total number of DC/NDC gold standard annotations.

POS	**# of *dA***	**POS**	**# of *dA***
NOUN	307	ADP	22
VERB	263	ADJ	22
PRON	70	NUM	7
CCONJ	67	AUX	3
ADV	59	DET	3
PRN	51	**Total**	874

Table 4: Distribution of grammatical classes that host *dA*.

4 The Machine Learning Approach

4.1 Features Used

We used a canonical set of features provided in Table 5 enabling a simple exploitation of local cues. Earlier work has revealed that the connective's syntactic context is a strong predictor of its DC role ((Lin et al., 2010; Gopalan and Lalitha Devi, 2016), among others). Semantic associations as simple as lexical cohesion manifested on surface through repeating words, or links inferred among propositions pose the harder problem in disambiguation. We observed that these hold for our data as well. The former is addressed by a reduction to parts of speech. The latter is crudely approximated via forms, assuming word-level selection is a signal for relevant constraints. Given that Turkish is a highly inflectional and agglutinating language, morphological variation entails the risk of leading the models to overlearn the declensions over a shared word root and result in the misclassification of *dA*, primarily because semantic information encapsulated in affixes is often tangential to our scope. To alleviate this potential noise, we implemented lemmatization and proceeded with the root forms. Finally, we integrated proper nouns to the feature set so as to capture cases like (2).

To model the context of *dA*, a discrete window size is defined according to standard locality and symmetry assumptions. Preliminary experiments hinted at an inverse relation between performance and text span, outputting a range of (-3, +3). Each line in Table 5 shows how we modeled the relation of three different features with *dA*'s context.

In Table 6, we illustrate what a data point looks like by showing various representation levels of example (2) above (see Table 7 for English glosses).

Features	Range	Definition
POS	(-3,+3)	The POS tags of 3 words before and 3 words after *dA*
LEMMA	(-3,+3)	The lemmas of 3 words before and 3 words after *dA*
ISPROPER	(-3,-1)	Whether one of the 3 words before *dA* is a proper noun or not

Table 5: The feature set for the usage disambiguation of *dA*.

Level	Form
I	Halil'in geldiğini fark etmediler. Halil **de** kadınlara bir şaka yapmaya karar verdi.
II	[*Halil'in geldiğini* *fark* *etmediler*]$_{a}$ [*Halil* **dA**$_{DC}$ *kadınlara* *bir* *şaka* *yapmaya karar verdi*]$_{b}$
III	FARK$_{(-3,N)}$ ET$_{(-2,V)}$ HALIL$_{(-1,N)}$ KADIN$_{(+1,N)}$ BIR$_{(+2,DET)}$ ŞAKA$_{(+3,N)}$

Table 6: A demonstrative example of three levels of representation of the data, namely, the *raw*, *annotated* and *encoded* levels of example (2). Level III is a projection of II onto a [-3,+3] window of *dA*'s immediate context. The boxed words in II correspond to the respective tokens in III. Each token is further lemmatized and tagged with POS information, resulting in the forms exploited by the learning model. The tags N, V, DET stand for noun, verb, determiner, respectively. The ISPROPER feature is excluded here for the sake of simplicity.

Level	Form
I	They didn't notice Halil came, **and** Halil decided to play a joke on the ladies.
II	[*They didn't notice* *Halil* *came*]$_{a}$ **and**$_{DC}$ [*Halil* *decided to play* *a* *joke* *on the* *ladies*]$_{b}$
III	NOTICE$^{i}_{(-3,N)}$ NOTICE$^{ii}_{(-2,V)}$ HALIL$_{(-1,N)}$ LADY$_{(+1,N)}$ A$_{(+2,DET)}$ JOKE$_{(+3,N)}$

Table 7: English glosses of the example demonstrated in Table 6. Note that *fark et* (*Eng.* 'notice') is a compound verb and taken as two parts tokenwise in the final step, which is denoted by superscripts (*i, ii*) in the gloss.

4.2 Experiments and Results

The *dA* corpus was processed before running ML algorithms over it. Firstly, since the number of DC and NDC samples in the corpus were not evenly distributed (cf. Table 3), we ran a few tests, and noticed a slight performance bias towards the more populated class. So, we pseudo-randomly excluded 72 DC samples and conducted the experiments on 802 data points (401 DC, 401 NDC).

The raw excerpts were processed by the UDPipe 2.0 pipeline (Straka et al., 2016; Straka and Straková, 2017) to obtain tagged and lemmatized discourse segments.

After constructing the final representations over POS tags, lemmas and proper nouns, on the transformed data, three supervised binary classifier models are trained based on Logistic Regression (LogRes) (Fan et al., 2008), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), and Random Forest (RF) (Ho, 1995) algorithms for comparison.

We used the *scikit-learn* (Pedregosa et al., 2011) library in a Python environment, and designed the sessions in 10 epochs all including 5-fold cross validation (CV), as standard test set evaluations may not be consistent about the impact of features when the characteristic variation throughout data is considered (e.g. Shi and Demberg, 2017). Each epoch contains 5 cycles shifting between 5 static slices on shuffled (randomly reindexed) data with 1-4, or 20%-80% test-train allocations. Then, the performance rate of the models is calculated by using the standard *classification report* and *confusion matrix* functions to obtain accuracy scores.

The models correctly disambiguated *dA* with the average accuracy of 0.77. Table 8 shows how performance oscillated across CV-cycles and epochs.[10]

4.3 Important Features for Classification

We trained each of our models to examine the predictive strength of each feature (and feature group) we used. Table 9[11] shows that POS is the most

[10] The highest scores achieved are written in **bold**.

[11] *lem* and *prn* are the abbreviations of *lemma* and *proper noun*, respectively.

Parameters	LogRes	SVM	RF
max. cycle	**0.82**	0.80	0.80
min. cycle	0.71	0.70	**0.70**
max. epoch	**0.79**	0.77	0.77
min. epoch	0.75	**0.74**	0.76
sd. (σ)	0.030	0.029	**0.028**
average	**0.77**	0.76	**0.77**

Table 8: Standard deviation, minimum, maximum, and average accuracy rates for classification with 5-fold CV in 10 epoch evaluation.

predictive feature solely achieving a minimum accuracy of 0.76. With all the features combined, the model reached an accuracy of 0.82 in the best case.

Features	LogRes	SVM	RF
pos+lem+prn	**0.82**	0.80	0.80
pos+lem	0.76	0.76	0.77
pos+prn	0.77	0.77	0.76
pos	0.76	0.77	0.76
lem+prn	0.73	0.73	0.74
lem	0.72	0.71	0.71

Table 9: Accuracy of the individual features used in the classification and the best combination.

4.4 Error Analysis

After calculating the success rates, we carried out an analysis to understand the possible causes of classification errors.

The major cause of classification errors is due to wrong POS tag assignment. This either happens when lemmatization is wrong or when the part-of-speech tagger fails to recognize noun-based (nominal or adjectival) predicates. For example, in (3), the verb *bekle* 'wait' at the -1 position, is wrongly lemmatized as 'bek' and assigned NOUN instead of VERB. For our models, being VERB at -1 is an important factor for the DC role of *dA* (e.g. (3), (7)), and mislabeling leads to an error in disambiguation. Logistic Regression and Random Forest sometimes correctly classify such tokens as DC, while SVM has not classified them as DC in any epoch. Hence, the false negative count increases.

Secondly, Turkish has nominal/adjectival predicates (sentences that do not contain an overt verb or auxiliary) such as the following:

(12) Ahmet doktor.
Ahmet doctor
'Ahmet is a doctor.'

Only having access to the surface form of a word, the part-of-speech tagger does not recognize the predicatehood of words like *güzel* '[is] nice' in (4) or *doktor* '[is] a doctor' in (12). These words are straightforwardly labelled as ADJ and NOUN, leading to mislabeling of the discourse connective usage of *dA* (also see Başıbüyük and Zeyrek (2023) for a detailed explanation of this kind of error).

5 Conclusion, Limitations and Further Work

Our work has two main parts; in the first part, we worked on a challenging annotation task not targeted before in Turkish NLP: the task of how the multi-faceted clitic *dA* can be annotated for its discourse and non-discourse connective usage. In the second part, we showed that with an ML approach, we can achieve success rates of an average of 0.77 in disambiguating the usage of *dA*.

However, our work is not without its limitations; for example, it is limited by the size of the corpus. It is assumed that as the dataset grows, more linguistic features of a discourse connective can be attested (Zeldes et al., 2019). Secondly, we are aware that the linguistic features we used in the ML experiments are not novel, but we believe we have shown that with a minimal set of rules, we can reach promising results in disambiguating the usage of *dA*.

Our work not only contributes to Turkish but also to discourse studies in general as we have brought to light the discourse role of a clitic through an annotation study and a computational analysis. It is therefore hoped to set the stage for other languages that have clitics with a discourse function. The results presented here can be used as a benchmark for Turkish clitics, and they can serve as a reference point for other languages that have clitics with a discourse function.

References

Uğurcan Arıkan, Onur Güngör, and Suzan Üsküdarli. 2019. Detecting clitics related orthographic errors in Turkish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 71–76, Varna, Bulgaria. INCOMA Ltd.

Nicholas Asher. 1993. *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

Kezban Başıbüyük and Deniz Zeyrek. 2023. Usage disambiguation of Turkish discourse connectives. *Language Resources and Evaluation*, 57(1):223–256.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Stefanie Dipper and Manfred Stede. 2006. Disambiguating potential connectives. In *Proceedings of KONVENS 2006 (Konferenz zur Verarbeitung natürlicher Sprache), Universität Konstanz*, pages 167–173, Konstanz.

Marcel Erdal. 2000. Clitics in Turkish. In *Studies on Turkish and Turkic languages : Proceedings of the Ninth International Conference on Turkish Linguistics*, pages 41–55, Wiesbaden. Harrassowitz.

Muharrem Ergin. 1975. *Türk Dil Bilgisi*. Bogazici Universitesi.

Martina Faller. 2020. The many functions of Cuzco Quechua =pas: implications for the semantic map of additivity. *Glossa: a journal of general linguistics*, 5(1):34.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(61):1871–1874.

Katherine Forbes-Riley, Bonnie Webber, and Aravind Joshi. 2006. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23(1):55–106.

Diana Forker. 2016. Toward a typology for additive markers. *Lingua*, 180:69–100.

Aslı Göksel and A Sumru Özsoy. 2003. *dA*: a focus/topic associated clitic in Turkish. *Lingua*, 113(11):1143–1167.

Sindhuja Gopalan and Sobha Lalitha Devi. 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 (BioTxtM2016)*, pages 89–98, Osaka, Japan. The COLING 2016 Organizing Committee.

Tin Kam Ho. 1995. Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, Montreal, QC, Canada.

Celia Kerslake. 1992. The role of connectives in discourse construction in Turkish. In *Modern studies in Turkish: Proceedings of the 6th International Conference on Turkish Linguistics*, pages 77–104, Eskişehir. Anadolu University, Education Faculty.

Ekkehard König. 2002. *The meaning of focus particles: A comparative perspective*. Routledge.

Majid Laali and Leila Kosseim. 2016. Automatic disambiguation of French discourse connectives. *International Journal of Computational Linguistics and Applications*, 7(1):11–30.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

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

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in Natural Language Processing*, pages 417–427, Berlin, Heidelberg. Springer Berlin Heidelberg.

Angela Schneider and Manfred Stede. 2012. Ambiguity in German connectives: A corpus study. In *Proceedings of KONVENS (Conference on Natural Language Processing) 2012*, pages 254–258.

Taner Sezer. 2017. TS Corpus project: An online Turkish dictionary and TS DIY Corpus. *European Journal of Language and Literature*, 3:18–24.

Taner Sezer and Bengü Sezer. 2013. TS Corpus: Herkes İçin Türkçe Derlem. In *Proceedings 27th National Linguistics Conference*, pages 217–225.

Wei Shi and Vera Demberg. 2017. On the need of cross validation for discourse relation classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 150–156, Valencia, Spain. Association for Computational Linguistics.

Yong-Siang Shih and Hsin-Hsi Chen. 2016. Detection, disambiguation and argument identification of discourse connectives in Chinese discourse parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1891–1902, Osaka, Japan. The COLING 2016 Organizing Committee.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, and Alan Lee. 2019a. Ambiguity in explicit discourse connectives. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 134–141, Gothenburg, Sweden. Association for Computational Linguistics.

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

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.

Deniz Zeyrek, Işın Demirşahin, AB Sevdik-Çallı, and Ruket Çakıcı. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse*, 4(2):174–184.

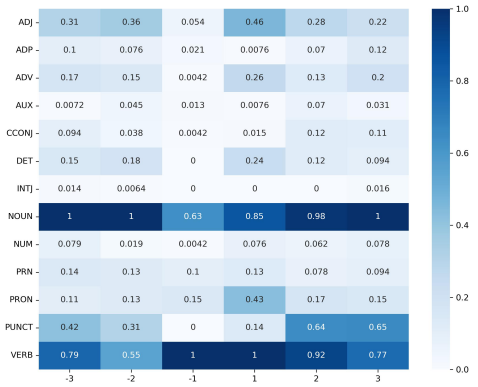
Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish Discourse Bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54(2):587–613.

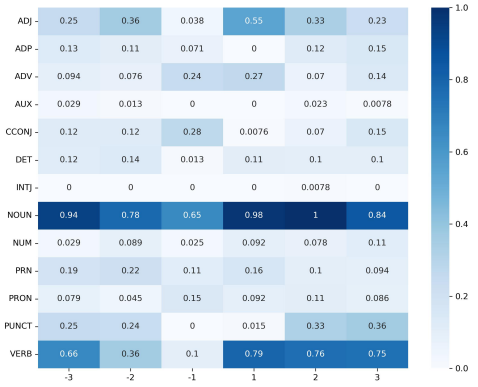
Arnold Zwicky. 1977. *On Clitics*. Bloomington: Indiana University Linguistics Club.

A Appendix

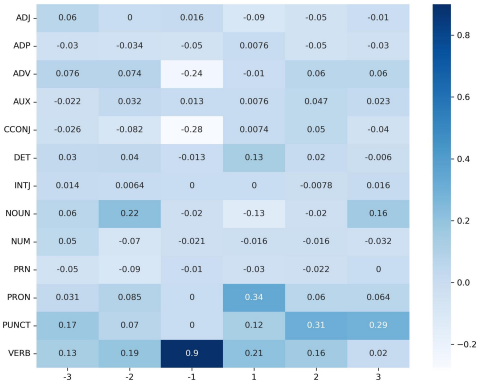
In the figures below, we present the distribution of POS types across DC/NDC instances.



(a)



(b)



(c)

Figure 1: Distribution of POS tags by their relative positions to *dA* within a [-3,+3] window, for DC and NDC samples (Figures 1*a* and 1*b*, respectively). Co-occurrence frequencies are scaled to [0-1] interval. Figure 1*c* represents the *rate of co-occurrence* difference between DC and NDC classes. Each value in the table satisfies the following condition: $c_{i,j} = a_{i,j} - b_{i,j}$. Convergence to 1 means a dominant DC characteristic at that specific position-POS correlation, and convergence to -1 means an NDC dominance.