



**HAL**  
open science

## Discourse parsing for multi-party chat dialogues

Stergos Afantenos, Eric Kow, Nicholas Asher, Jérémy Perret

► **To cite this version:**

Stergos Afantenos, Eric Kow, Nicholas Asher, Jérémy Perret. Discourse parsing for multi-party chat dialogues. Conference on Empirical Methods on Natural Language Processing (EMNLP 2015), Sep 2015, Lisbon, Portugal. pp. 928-937. hal-01535954

**HAL Id: hal-01535954**

**<https://hal.science/hal-01535954>**

Submitted on 9 Jun 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : <http://oatao.univ-toulouse.fr/>  
Eprints ID : 16912

The contribution was presented at EMNLP 2015 :  
<http://www.emnlp2015.org/>

**To cite this version** : Afantenos, Stergos and Kow, Eric and Asher, Nicholas and Perret, Jérémy *Discourse parsing for multi-party chat dialogues*. (2015) In: Conference on Empirical Methods on Natural Language Processing (EMNLP 2015), 17 September 2015 - 21 September 2015 (Lisbon, Portugal).

Any correspondence concerning this service should be sent to the repository administrator: [staff-oatao@listes-diff.inp-toulouse.fr](mailto:staff-oatao@listes-diff.inp-toulouse.fr)

# Discourse parsing for multi-party chat dialogues

Stergos Afantenos Eric Kow Nicholas Asher Jérémy Perret

IRIT, Université Toulouse & CNRS, Univ. Paul Sabatier,

118 Route de Narbonne, 31062 Toulouse

{firstname.lastname@irit.fr}

## Abstract

In this paper we present the first ever, to the best of our knowledge, discourse parser for multi-party chat dialogues. Discourse in multi-party dialogues dramatically differs from monologues since threaded conversations are commonplace rendering prediction of the discourse structure compelling. Moreover, the fact that our data come from chats renders the use of syntactic and lexical information useless since people take great liberties in expressing themselves lexically and syntactically. We use the dependency parsing paradigm as has been done in the past (Muller et al., 2012; Li et al., 2014). We learn local probability distributions and then use MST for decoding. We achieve 0.680  $F_1$  on unlabelled structures and 0.516  $F_1$  on fully labeled structures which is better than many state of the art systems for monologues, despite the inherent difficulties that multi-party chat dialogues have.

## 1 Introduction

Discourse parsing is a difficult, multifaceted problem involving the understanding and modeling of various semantic and pragmatic phenomena as well as understanding the structural properties that a discourse graph can have. Unsurprisingly, most extant theories and computational approaches postulate an extremely simplified version of discourse structure. One of the most widely cited theories, Rhetorical Structure Theory (RST) (Mann and Thompson, 1987; Mann and Thompson, 1988; Taboada and Mann, 2006), requires that only *adjacent* discourse units be connected together with a discourse relation. Another widely cited approach, the Penn Discourse Treebank (PDTB) (Prasad et

al., 2008), focuses on decisions about the discourse connectives that label the attachment of potentially arbitrary text spans but does not make any claims as to what the overall discourse structure of the resulting annotation looks like. Further, all computational work on the PDTB takes the attachments as given in discourse parsing tasks. In both cases, the attachment problem, finding which discourse units are attached to which, is vastly simplified, though this has enabled researchers to explore various approaches for discourse parsing (Marcu, 2000; Sagae, 2009; Hernault et al., 2010; Joty et al., 2012).

Our paper’s main contribution is to provide a discourse parsing model for multi-party chat dialogue (*i.e.* typed online dialogue), trained on a large corpus we have developed annotated with full discourse structures. We study attachment problem in detail for this genre, without using the simplifying hypotheses mentioned above that we know to be inadequate. In the following section, we describe the *Settlers of Catan* game and our corpus in more detail and discuss some problematic structures for discourse parsing from our corpus. We motivate our choice of a particular discourse theory, the Segmented Discourse Representation Theory (SDRT), as the underlying theoretical model for our annotations. In section 4 we present our parsing approach, which consists of building a local probability distribution model which serves as input to a series of decoder mechanisms. We present and discuss the results we obtain in section 5, while related work and conclusions are presented in sections 6 and 7 respectively.

## 2 What’s so special about multi party dialogue?

Multi-party dialogue or multi-party chat involves multiple interlocutors that may address one or more interlocutors during their turn. For example,

a person might pose a question that concerns all the participants; and once everybody has replied, that same person might reply to all of them with a single comment (e.g. thanking them) or with a single acknowledgment. Figure 1 provides an example from our corpus. In turn 234, gotwood4sheep asks a question and makes an underspecified offer to all the players. He then gets back negative responses to his question from inca, Cheshire-CatGrin and dmm; and then he broadcasts in 239 an acknowledgment of all the negative responses. That is, we have 235, 236 & 238 all attached to 234 as answers to the question in 234; and we have 239 that is attached to 235, 236 & 238 as an acknowledgment of the contents of those turns. A graphical representation is shown on the right of the same picture.

The presence of such structures makes a powerful case that the general framework guiding the annotation of multi-party dialogues should take *non-tree-like* graphs as the basic form of discourse structures. This will require then rethinking the task of discourse parsing when attempting to learn such structures. In particular, the following questions present themselves: 1) how many non-tree-like structures are there? 2) what are the constraints on discourse graphs, if they are not trees? 3) how far can traditional tree-based decoding mechanisms get us in dealing with such data?

Another complicated phenomenon in multi-party chat dialogues is the presence of crossing dependencies. Many theories of discourse structure like RST, given that they allow attachment only of adjacent spans will perforce not allow structures with crossing dependencies. Also theories that postulate a simple *right frontier constraint*, according to which only elements on the right frontier of a discourse structure (whether graph or tree) will in general not generate structures with crossing dependencies. However, crossing dependencies are commonplace in multi-party chat. Several subgroups of interlocutors can and do momentarily form and carry on a discussion amongst themselves, forming thus multiple concurrent discussion threads. Since, though, what is being written is publicly available to all involved parties, it can be the case that participants of one thread might reply or comment to something said to another thread. Figure 2 contains an example from our corpus.

There are at least three threads in this excerpt,

and we have given them different fonts to aid the reader. The intuitive attachments in this excerpt involve the following crossing dependencies: (165, 168), (167,170), (176, 178), (177, 179), (175, 181), (177,182), and (180,183). We note also the lack of standard discourse markers such as those found in the PDTB or RST manuals, “personalized” orthography, the lack of elaborate syntactic structure and the frequent presence of sentence fragments, all of which means we cannot rely on sentential syntax to aid with discourse parsing (syntax is very useful in monologue discourse parsing, as witnessed by the dramatically higher scores for intra-sentential discourse parsing (Joty et al., 2015)). Multi-party dialogue presents a discourse parsing problem free of syntactic crutches.

The phenomena we have just described are just some of the complications that appear in the discourse representation of multi-party dialogues, unfortunately rendering discourse theories based on attaching only adjacent units unsuitable for the representation of multi-party dialogues. In order to be able to capture the discourse phenomena present in our chat corpus, we have decided to use the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). This theory not only allows long distance attachments, which (Ginzburg, 2012) finds attested in multi-*logue*, but also has semantics capable of dealing with fragments or non sentential utterances (Schlangen, 2003), which are frequent in our corpus. Also, it can model non-tree like structures, like that shown in Figure 1, which account for at least 9% of the links in our corpus. Such structures make theories that model discourse structures with rooted trees, like *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1987) or simple dialogue models where attachments are always made to *Last*—cf. (Schegloff, 2007; Poesio and Traum, 1997)—unsuitable.

A final feature of discourse annotations that multi-party dialogue and monologue share is the presence of complex discourse units or CDUs. CDUs are in fact subgraphs of the discourse graph that have a rhetorical function or bear some discourse relation to another constituent. Examples are easy to come by. Consider the following example:

- (1) gw: Do you have a sheep?  
th: I do, if you give me an ore

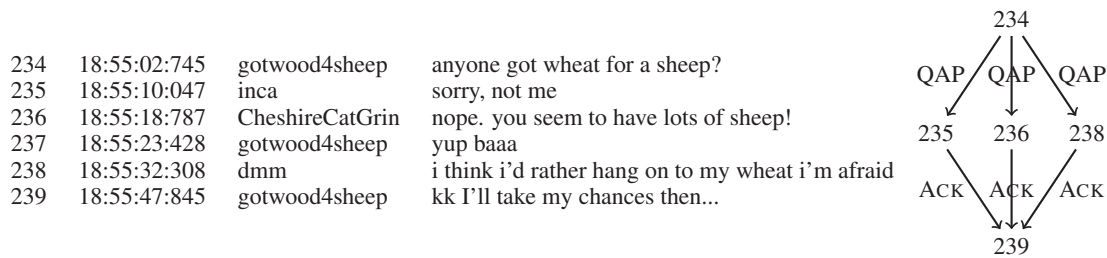


Figure 1: Dialogue excerpt showing the need for general graphs instead of trees.

165	lj	<b>anyone want sheep for clay?</b>
166	gw	<b>got none, sorry :(</b>
167	gw	so how do people know about the league?
168	wm	<b>no</b>
170	lj	i did the trials
174	tk	i know about it from my gf
175	gw	[yeah me too,] <sub>a</sub> [are you an Informatics student then, lj?] <sub>b</sub>
176	tk	did not do the trials
177	wm	<b>has anyone got wood for me?</b>
178	gw	[I did them] <sub>a</sub> [because a friend did] <sub>b</sub>
179	gw	<b>lol william, you cad</b>
180	gw	<b>afraid not :(</b>
181	lj	<i>no, I'm about to start math</i>
182	tk	<b>sry no</b>
183	gw	<b>my single wood is precious</b>
184	wm	<b>what's a cad?</b>

Figure 2: Example of interleaved threads

th: or a wood.

Clearly, th's two turns combine to form a CDU that is then related by a conditional discourse relation to *I do*. That is *you give me an ore* and *or a wood* form together the antecedent to the conditional that he expresses. In order to reflect this semantic dependency, SDRT creates collections of Elementary Discourse Units (EDUs) forming a coherent discourse unit (called Complex Discourse Unit, CDU) and link it to any other discourse unit. The end result of this process is the creation of a hypergraph or, equivalently, a graph with two types of edges. Thus, our general conception of a discourse structure for a discourse  $D = \{e_1, \dots, e_n\}$ , where  $e_i$  are the EDUs of  $D$ , is a tuple  $(V, E_1, E_2, \ell)$ , where  $V$  is a set of nodes or discourse units including  $\{e_1, \dots, e_n\}$ ,  $E_1 \subseteq V \times V$  a set of edges representing discourse relations and  $E_2 \subseteq V \times V$  a set of edges that represents parthood in the sense that if  $(x, y) \in E_2$ , then  $x$  is a discourse unit that is an element of the CDU  $y$ .  $\ell: E_1 \rightarrow \text{Relations}$  is a function that assigns each arc a discourse relation type. Our corpus contains many instances of CDUs, some of which are quite large, encompassing an entire question answering session like that

seen in Figure 1.

### 3 The STAC corpus

The corpus that we use was collected from an on-line version of the game *The Settlers of Catan*. *Settlers* is a multi-party, win-lose game in which players use resources such as wood and sheep to build roads and settlements. In the standard on-line version, players interact solely through the game interface, making trades and building roads, etc., without saying a word. In our online version, players were asked to discuss and negotiate their trades via a chat interface before finalizing them non-linguistically via the game interface. As a result, players frequently chatted not only to negotiate trades, but to discuss numerous topics, some unrelated to the task at hand.

The *Settlers* corpus is ideal for studying multilogue. First, while the chats maintain the advantage of written text (no need for transcription), they approximate spoken communication. We have to deal with many sentence fragments, non-standard orthography and sometimes lack of syntax. Second, they manifest phenomena particular to multilogue, such as multiple conversation threads and non-tree-like structures.

The corpus consists of 59 games out of which 36 games have so far been annotated for discourse structure in the style of SDRT. Each game consists of several dialogues representing a single turn of the game. Each dialogue is treated as a separate document. About 10% of our corpus was held out for evaluation purposes while the rest was kept for training. Detailed statistics on the number of dialogues, EDUs and relations contained in each sub-corpus can be found in table 1.

The dialogues in our corpus have an average size of 10 EDUs with 8 speaker turns, though the longest has 156 EDUs and 119 turns. The vast majority of our discourse connections thus lie between turns. All dialogues also have a dialogue act

	Total	Training	Testing
Dialogues	1091	968	123
EDUs	10677	9545	1132
Relations	11348	10158	1190

Table 1: Dataset overview

style annotation in which each EDU is assigned a particular type (it can be an offer or counter-offer, an acceptance or refusal, or other) (Sidner, 1994a; Sidner, 1994b). The dialogue act annotations have been used to train an automatic classifier for EDUs (Cadilhac et al., 2013). This large annotation effort was carried out by 4 annotators who had no special knowledge of linguistics, but who received training over 22 negotiation dialogues with 560 turns. Because annotating full discourse structures is a very complex task (using an exact match criterion of success, the inter annotator agreement score was a Kappa of 0.72 attachment on structures, 0.58 on labelling (Afantenos et al., 2012a)), experts made several passes over the annotations from the naive annotators, improving the data and debugging it.

The discourse graphs in our development corpus exhibited several interesting properties. First of all they are DAGs with a unique root, one unit that has no incoming edges. Secondly, the graphs are weakly connected in almost all cases: i.e., every discourse unit in it is connected to some other discourse unit. Thirdly, our graphs are *reactive* in the sense that speakers’ contributions are reactions and attach anaphorically to prior contributions of other speakers. This means that edges between the contributions of different speakers are always oriented in one direction. This is a general feature of dialogue, as we explain in the next section.

## 4 Model

### 4.1 Dependency Structures

For a given discourse graph for SDRT of the form  $(V, E_1, E_2, \ell)$ , we have as yet no general and reliable method to calculate edges in  $E_2$ ; and no such method has been presented in the literature. In order to perform constrained decoding over local probability distributions, we have opted for a strategy first presented in Muller et al. (2012) for SDRT. The strategy involves transforming hypergraphs into dependency graphs. We transform our full graphs  $(V, E_1, E_2, \ell)$  into dependency struc-

tures  $(V', E_1, \ell)$ , with  $V' \subset V$  the set of EDUs in  $V$  by replacing any attachment to a CDU with an attachment to the CDU’s head—the textually first EDU within the CDU which has no incoming links. Our transformation in effect sets  $E_2$  in our general definition of a graph to  $\emptyset$ . In the case that we have a discourse relation between two EDUs, this relation is kept intact since it already represents a dependency arc. In case a discourse relation has one or two CDUs as arguments, the CDUs need to be replaced with their *recursive head*. In order to calculate the recursive head we identify all the DUs with no incoming links; if they are CDUs we recursively apply the algorithm until we get an EDU. If there is more than one EDU with no incoming links we pick the leftmost, i.e. the one firstly introduced in the text.

Hirao et al. (2013) and Li et al. (2014) later followed a similar strategy for the creation of dependency structures for RST. Every single nucleus-satellite relation was transformed into a dependency relation with the governor being the EDU representing the nucleus and the dependent being the satellite. For relations between non-EDU higher spans, the recursive head was used. It is unclear how Li et al. (2014) deal with binary multi-nucleus relations like CONTRAST for example; it is not clear how to calculate the recursive head of the span.<sup>1</sup> In such cases an arbitrary decision—like always taking as the nucleus the leftmost or the rightmost span—has to be taken. In the SDRT annotations, however, every edge in the graph is already directed and so such arbitrary decisions can be avoided.

Ideally, what one then wants is to learn a function

$$h : \mathcal{X}_{E^n} \mapsto \mathcal{Y}_{\mathcal{G}}$$

where  $\mathcal{X}_{E^n}$  is the domain of instances representing a collection of EDUs for each dialogue and  $\mathcal{Y}_{\mathcal{G}}$  is the set of all possible SDRT graphs. However, given the complexity of this task and the fact that it would require an amount of training data that we currently lack in the community, we aim at the more modest goal of learning a function

$$h : \mathcal{X}_{E^2} \mapsto \mathcal{Y}_R$$

where the domain of instances  $\mathcal{X}_{E^2}$  represents features for a pair of EDUs and  $\mathcal{Y}_R$  represents the

<sup>1</sup>Although Li et al. (2014) do explain how to treat n-ary multinuclear relations, following others (Hernault et al., 2010, for example).

set of SDRT relations. The upshot of this is that we are building a local sort of model that learns relations between individual EDUs with a certain probability but that it does not learn a local or even global structure.

One of the drawbacks of this approach, however, is that it does not guarantee an object that is well formed. Learning a probability distribution over EDUs and then choosing the most probable relation or attachment for each pair of EDUs potentially leads to structures that contain cycles. To avoid this, we can't blindly choose the most probable relation or attachment decision for each pair of EDUs. Instead, we should use this probability distribution as an input to a decoding mechanism.

## 4.2 Local probability distributions

We used a regularized maximum entropy (shortened as MaxEnt) model (Berger et al., 1996). In MaxEnt, we estimate the parameters of an exponential model of the following form:

$$P(r|p) = \frac{1}{Z(c)} \exp \left( \sum_{i=1}^m w_i f_i(p, r) \right)$$

where  $p$  represents the current pair of EDUs and  $r$  the learnt label (i.e. the type of relation, or a binary attachment value between the two EDUs). Each pair of EDUs  $p$  is encoded as a vector of  $m$  indicator features  $f_i$  (see table 2 for more details). There is one weight/parameter  $w_i$  for each feature  $f_i$  that predicts its classification behavior. Finally,  $Z(c)$  is a normalization factor over the different class labels, which guarantees that the model outputs probabilities. In MaxEnt, the values for the different parameters  $\hat{w}$  are obtained by maximizing the log-likelihood of the training data  $T$  with respect to the model (Berger et al., 1996):

$$\hat{w} = \operatorname{argmax}_w \sum_i^T \log P(r^{(i)}|p^{(i)})$$

## 4.3 The turn constraint

Given our observations about the structure of dialogues in our corpus, we hypothesize that a dialogue is fundamentally sequential: first one person talks and then others react to them or ignore them, but the discourse links that do occur between speaker turns are *reactive*. In other words, a turn can't be anaphorically and rhetorically dependent on a turn that comes after it. Thus, the

nature of dialogue imposes an essential and important constraint on the attachment process that is not present for monologue or single-authored text, where an EDU may be dependent upon any EDU, later in the ordering or not: in dialogue there are no "backwards" rhetorical links such that an EDU in turn  $n$  by speaker  $a$  is rhetorically and anaphorically dependent upon an EDU in turn  $n + m$  of speaker  $b$  with  $a \neq b$ . We call this the *Turn Constraint*. Within a turn, however, just as in monologue (as is evident from a study of most styles of discourse annotations of text), backwards links are allowed.

Given this observation, we decided to split our local model into two different ones. The first one concerns the learning of a model for intra-turn utterances,<sup>2</sup> while the second models inter-turn utterances. The intra-turn model considers as input during learning all pairs of EDUs  $(i, j)$  with  $i \neq j$ . The inter-turn model on the other hand does not contain any backward links during learning. In other words it takes as input all pairs of EDUs  $(i, j)$  with  $i < j$ . We apply the turn constraint not only during learning of the local models, but also during decoding. This practice is also followed—at the sentence level—for monologues (Wellner and Pustejovsky, 2007; Joty et al., 2012; Joty et al., 2013), though our turn constraint, we believe, is firmly supported not only by our data but also by a good theoretical model of dialogue.

## 4.4 Decoders

The local probability distributions obtained are used as decoder inputs. We have experimented with several decoders. As a baseline measure we have included what we call a LOCAL decoder which creates a simple classifier out of the raw local probability distribution. In the case of MaxEnt, for example, this decoder selects

$$\hat{r} = \operatorname{argmax}_r \left( \frac{1}{Z(c)} \exp \left( \sum_{i=1}^m w_i f_i(p, r) \right) \right)$$

with  $r$  representing a relation type or a binary attachment value. We also used the baseline LAST, where each EDU is attached to the immediately preceding EDU in the linear, textual order.

<sup>2</sup>EDUs are considered as belonging to the same turn if they are by the same speaker without any interjection from an other speaker. In other words any consecutive EDU by the same speaker is considered as belonging to the same turn.

**Maximum Spanning Trees** To answer our questions, “how many non-tree-like structures are there?” and “how far can tree decoding algorithms get us in multi-party dialogue?”, our first decoder is the classic Maximum Spanning Trees (MST) algorithm— used by McDonald et al. (2005) for syntactic dependency parsing as well as by Muller et al. (2012) and Li et al. (2014) for discourse parsing—tweaking it in order to produce structures that are closer to the ones specific to multi-party dialogue. We are looking for:

$$T^* = \underset{T \text{ a spanning tree of } G}{\operatorname{argmax}} \sum_{e \in E(T)} w(e)$$

$$w(e) = \log \left( \frac{p(e)}{1 - p(e)} \right)$$

$G$  being the complete graph of possible edges returned by the classifiers ;  $E(D)$  representing the edges of  $D$ . The weight function  $w$  computes the log-odds of the probability returned by the model.

We used Chu-Liu-Edmonds version of the MST algorithm (Chu and Liu, 1965; Edmonds, 1967), which requires a specific node to be the root, i.e. a node without any incoming edges, of the initial complete graph. For each dialogue, we made an artificial node as the root with special dummy features. At the end of the procedure, this node points the real root of the discourse graph.

**Combining intra- and inter-turn models with the turn constraint** As described above, we learn a separate local model for intra- and inter-turn EDUs. We also use the turn constraint during decoding. For the intra-turn decoding, we have experimented with various options. One concerns the creation of a classifier out of the local probability distribution. Another intra-turn decoder is Last, which always takes the last EDU for attachment. Finally we also used MST.

We used the exact same decoding approaches for inter-turn decoding. With the structure for the inter-turn EDUs produced separately, we replace those structures with their heads. The detection of a structure’s head, be it intra- or inter-turn, uses the same trick as McDonald et al. (2005) did for syntactic parsing: inserting a dummy node as a fake head which contains only outgoing links enabling us essentially to learn the real head of our structures. Our best overall model used Last to link EDUs inside the turn together with MST and the turn constraint for predicting the global structure.

Category	Description
Positional	Speaker initiated the dialogue
-	First utterance of the speaker in the dialogue
-	Position in dialogue
-	<i>Distance between EDUs</i>
-	<i>EDUs have the same speaker</i>
Lexical	Ends with exclamation mark
-	Ends with interrogation mark
-	Contains possessive pronouns
-	Contains modal modifiers
-	Contains words in lexicons
-	Contains question words
-	Contains a player’s name
-	Contains emoticons
-	First and last words
Parsing	Subject lemmas given by syntactic dependency parsing
-	Dialogue act according to predicting model

Table 2: Feature set description. Pair features are italicized.

## 5 Experiments and Results

To train our local models, we extracted features for every pair of EDUs in a given dialogue. Our features concern the pair of EDUs as well as features related to each EDU specifically. The feature set, detailed in Table 2, can be summarized as follows:

- Positional features: (related to) the non-linguistic context of the pair;
- Lexical features: single words<sup>3</sup> and punctuation present in the EDUs;
- Parsing features: dependency<sup>4</sup> and dialogue act<sup>5</sup> tagging.

Table 3 shows our results on our unseen test corpus, which contains a randomly selected 10% of dialogues in our corpus. The best configuration was selected after performing ten-fold cross validation on the training corpus. The reported results implement the turn-constraint during training

<sup>3</sup>We use a number of lexicons (opinion markers, quantifiers, etc.), each corresponding to a feature

<sup>4</sup>Provided by the Stanford CoreNLP pipeline (Manning et al., 2014).

<sup>5</sup>The prediction model of Cadilhac et al. (2013) generates EDU tags such as *Offer*, *Refusal*, etc.



for the local models. In other words, training instances for the local models include only forward links.

We used two baselines. The first one, Last, simply attaches every EDU to its previous one. This is a very strong baseline in discourse parsing (Muller et al., 2012, for example). The second baseline is essentially the local classifier without any further decoding; in other words, we simply select the class with the highest probability both for attachment and labeling. Attaching to last gives us an F-score of 0.584 for attachment and 0.391 when we add the relations as well. Using only classification from the local probability distribution without decoding gives 0.541 for attachment and 0.446 for attachments and relations.

The best results for the global parsing problem exploited the turn constraint both during learning the local model and during decoding. Within a turn, our discourse structures are simple and largely linear; the best intra-turn results came from using Last. Most of our interlocutors did not create elaborate discourse structures with long-distance attachments within the same turn. The inter-turn level was a different story, as the figures show. For inter-turn and the global problem, MST using the heads of the intra-turn substructures computed with Last, produced the best results. The F1 score for unlabeled structures is at 0.671 while for labelled structures we have 0.516. To enable a comparison with RST style parsing where exact arguments for discourse relations are not computed, the undirected attachment F1 score = 0.68 for the global parsing problem.

Despite the inherent difficulty of discourse parsing on multi-party chat dialogues (simultaneous, multiple discussion threads, lack of syntax) our results are close to or better than the current state of the art for discourse parsing on monologue. There are two approaches currently that use dependency parsing strategies for discourse, thoroughly described in the next section. Li et al. (2014) report an accuracy of 0.7506 for unlabelled structures and 0.4309 for the full labelled structures. Muller et al. (2012) report 0.662 for unlabelled structure and 0.361 for labelled structures. We outperform both systems for full labelled structures, and despite our non-tree-like structures beat or are close to these on unlabelled attachments. Though comparisons across different corpora are difficult, the numbers suggest that our results are more than

competitive. Our results also suggest that one can get quite far with tree-based decoding algorithms, though we know that in principle MST cannot do better than 91% even with a perfect local model (a model in which an arc is giving probability 1 just in case it occurs in the gold standard annotation).

## 6 Related Work

To date, discourse parsing has almost exclusively been applied to monologue. Multi-party chat dialogues have never been considered before. Baldridge and Lascarides (2005) predicted tree discourse structures for 2 party “directed” dialogues from the Verbmobil corpus by training a PCFG that exploited the structure of the underlying task. Elsner and Charniak (2010), Elsner and Charniak (2011) are presenting a combination of local coherence models initially provided for monologues showing that those models can satisfactorily model local coherence in chat dialogues. Nonetheless they do not present a full-fledged discourse parsing model. Our data required a more open domain approach and a more sophisticated approach to structure.

Our use of dependency parsing for learning discourse structure has a few antecedents in the literature on monologue. One of the first papers to introduce this technique is Muller et al. (2012), working with a small French language corpus, ANNOTODIS (Afantenos et al., 2012a). They use a similar approach to us, including the classic version of the MST decoder. They also used an  $A^*$  search as another decoding mechanism but it gave the same results as MST. As we have said, our results better theirs both on attachment and full labeled structures. In the context of RST, Hirao et al. (2013) and Li et al. (2014) transform RST trees into dependency structures; we have discussed their in section 4.1. Li et al. (2014) use both the Eisner algorithm (Eisner, 1996) as well as the MST algorithm from McDonald et al. (2005). As we mentioned, our labelled scores are higher than theirs, though we are cautious of making comparisons across such different corpora.

Most work on discourse parsing focuses on the task of discourse relation labeling between pairs of discourse units—e.g., Marcu and Echihiabi (2002) Sporleder and Lascarides (2005) and Lin et al. (2009). This corresponds to our local model. As we have shown in this paper, this setting makes an unwarranted assumption, as it assumes inde-

Method	Undirected Attachment			Directed Attachment			Full Labelled Structure		
	prec	rec	F1	prec	rec	F1	prec	rec	F1
LAST	0.602	0.566	0.584	0.602	0.566	0.584	0.403	0.379	0.391
LOCAL	0.698	0.488	0.574	0.623	0.478	0.541	0.513	0.394	0.446
INTRA-TURN	0.837	0.955	0.892	0.808	0.922	0.861	0.489	0.558	0.521
INTER-TURN	0.617	0.516	0.562	0.616	0.514	0.561	0.492	0.411	0.448
GLOBAL	0.697	0.663	0.680	0.688	0.655	0.671	0.529	0.503	0.516

Table 3: Evaluation results.

pendence of local attachment decisions. There is also work on discourse structure within a single sentence; e.g., Soricut and Marcu (2003) makes use of dynamic programming along with a standard bottom-up chart parsing, while Sagae (2009) uses shift-reduce algorithm for intra-sentential discourse analysis. Such approaches do not apply to our data, as most of the structure in our dialogues lies beyond the sentence level.

As for document-level discourse parsers, Subba and Di Eugenio (2009) use a transition-based approach, following the paradigm of Sagae (2009). duVerle and Prendinger (2009) and Hernault et al. (2010) both rely on locally greedy methods. Like us, they treat attachment prediction and relation label prediction as independent problems. Feng and Hirst (2012) extend this approach by additional feature engineering but is restricted to sentence-level parsing. Finally, Joty et al. (2012) present a sentence-level discourse parser that uses Conditional Random Fields to capture label interdependencies and chart parsing for decoding. Joty et al. (2013) and Joty et al. (2015) extend this approach on the level of documents and have the best results non-dependency based discourse parsing, with an F1 of 0.689 on unlabelled structures and 0.5587 on labelled structures. Our scores are very close to Joty et al.’s, however, and achieved with much simpler methods than theirs.

## 7 Conclusions

As far as we know, this is the first paper to deal with discourse parsing in multi-party chat dialogues. We believe that such data will be useful for other discourse parsing tasks like analyzing fora with multi-threads. We have used the STAC corpus (Afantenos et al., 2012b) for our data. To simplify the parsing task, we transformed our SDRT structures into dependency ones. We used two different local probability distribution models as input to several decoding mechanisms, including one based on the Maximum Spanning Tree al-

gorithm, and an enhanced version of it in order to produce structures closer to the ones we observe. We obtain the best results using the enhanced version of the MST algorithm. In future work, we plan to investigate ILP constraints in greater depth to develop a plausible alternative to MST on DAGs.

## References

- Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cecile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paul Pery-Woodley, Laurent Prevot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012a. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cédric Degremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Verena Rieser, and Laure Vieu. 2012b. Developing a corpus of strategic conversation in the settlers of catan. In Noriko Tomuro and Jose Zagal, editors, *Workshop on Games and NLP (GAMNLP-12)*, Kanazawa, Japan.
- N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.
- Jason Baldridge and Alex Lascarides. 2005. Probabilistic head-driven parsing for discourse structure. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)*.
- A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.
- Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict

- trades in a win-lose game. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 357–368, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Y. J. Chu and T. H. Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400.
- David duVerle and Helmut Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 665–673, Suntec, Singapore, August. Association for Computational Linguistics.
- Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71B(233–240).
- Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, volume 1, pages 340–345, Copenhagen, Denmark.
- Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*, 36(3):389–409.
- Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1179–1189, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68, Jeju Island, Korea, July. Association for Computational Linguistics.
- Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.
- Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3):1–33.
- Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. In *EMNLP-CoNLL*.
- Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*.
- Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland, June. Association for Computational Linguistics.
- Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore, August. Association for Computational Linguistics.
- William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A Framework for the Analysis of Texts. Technical Report ISI/RS-87-185, Information Sciences Institute, Marina del Rey, California.
- William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8(3):243–281.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL*, pages 368–375.
- Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.
- Ryan T. McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT/EMNLP*.
- Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India,

- December. The COLING 2012 Organizing Committee.
- Massimo Poesio and David R Traum. 1997. Conversational actions and discourse situations. *Computational intelligence*, 13(3):309–347.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*.
- Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies, IWPT '09*, pages 81–84, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Emanuel A Schegloff. 2007. *Sequence organization in interaction: Volume 1: A primer in conversation analysis*, volume 1. Cambridge University Press.
- David Schlangen. 2003. *A coherence-based approach to the interpretation of non-sentential utterances in dialogue*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.
- C. Sidner. 1994a. An artificial discourse language for collaborative negotiation. In *AAAI*, volume 1, pages 814–819. MIT Press, Cambridge.
- C. Sidner. 1994b. Negotiation in collaborative activity: a discourse analysis. *Knowledge-Based Systems*, 7(4):265 – 267.
- R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 149–156. Association for Computational Linguistics.
- Caroline Sporleder and Alex Lascarides. 2005. Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Bulgaria.
- Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574, Boulder, Colorado, June. Association for Computational Linguistics.
- Maite Taboada and William C. Mann. 2006. Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse Studies*, 8(3):423–459, June.
- Ben Wellner and James Pustejovsky. 2007. Automatically Identifying the Arguments of Discourse Connectives. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language*

*Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 92–101, Prague, Czech Republic, June. Association for Computational Linguistics.