Dialogue & Discourse 8(1) (2017) 106–131 doi: 10.5087/dad.2017.104

A Psycholinguistic Model for the Marking of Discourse Relations

Frances Yung PIKYUFRANCES-Y@IS.NAIST.JP
Computational Linguistics, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, Japan

Kevin Duh KEVINDUH@CS.JHU.EDU
Human Language Technology Center of Excellence, John Hopkins University
Stieff Building, 810 Wyman Park Drive, Baltimore, MD 21211-2840, USA

Taku Komura TKOMURA@INF.ED.AC.UK
Institute of Perception, Action and Behaviour, School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom

Yuji Matsumoto MATSU@IS.NAIST.JP
Computational Linguistics, Nara Institute of Science and Technology
8916-5 Takayama Ikoma, Nara, Japan

Editor: Maite Taboada

Submitted 5/2016; Accepted 12/2016; Published online 1/2017

Abstract

Discourse relations can either be explicitly marked by discourse connectives (DCs), such as *therefore* and *but*, or implicitly conveyed in natural language utterances. How speakers choose between the two options is a question that is not well understood. In this study, we propose a psycholinguistic model that predicts whether or not speakers will produce an explicit marker given the discourse relation they wish to express. Our model is based on two information-theoretic frameworks: (1) the Rational Speech Acts model, which models the pragmatic interaction between language production and interpretation by Bayesian inference, and (2) the Uniform Information Density theory, which advocates that speakers adjust linguistic redundancy to maintain a uniform rate of information transmission. Specifically, our model quantifies the *utility* of using or omitting a DC based on the expected surprisal of comprehension, cost of production, and availability of other signals in the rest of the utterance. Experiments based on the Penn Discourse Treebank show that our approach outperforms the state-of-the-art performance at predicting the presence of DCs (Patterson and Kehler, 2013), in addition to giving an explanatory account of the speaker's choice.

Keywords: discourse connectives, psycholinguistics, rational speech acts model, uniform information density

1. Introduction

Speakers or authors produce informative utterances such that listeners or readers can understand the intended message.[1] Grice's *Maxim of Quantity* states that human speakers communicate by being as informative as required, but no more (Grice, 1975). If a speaker always tries to provide as much information as possible, the resulting utterance could become excessively long and tedious. Such utterance not only takes effort for the speaker to produce, but also contains redundant information that is not necessary for the listener.

In this work, we model how speakers optimally plan the presentation of discourse structure in terms of informativeness. Specifically, we propose a model that predicts whether speakers will use or omit a discourse connective, given the sense of the discourse relation they want to convey. Discourse relations are relations between unit of texts (known as *arguments*) that make a document coherent. These relations can be explicitly marked in the surface text or inferred by the readers, as shown in Examples 1a to 1c.

(1a) It was a great movie, **but** I did not like it.

(1b) It was a great movie, **therefore** I liked it.

(1c) It was a great movie. I liked it.

In Example 1a, the word *but* indicates a CONTRAST relation, and *therefore* indicates a RESULT relation in Example 1b. We call *but* and *therefore explicit* discourse connectives (DCs). In Example 1c, DCs are absent, but a RESULT relation can be inferred. We say the two sentences (called *arguments*) are connected by an implicit DC.

Explicit and implicit relations differ in their level of ambiguity. Explicit relations can be signaled by a variety of lexical, syntactic and semantic features, of which DCs are the most informative cues to identify discourse relations (Pitler et al., 2008).[2] In contrast to explicit relations, implicit relations are more ambiguous. For example, *I liked it* can also be read as a JUSTIFICATION for the first sentence in Example 1c.

However, marking a discourse relation or not is subject to not only ambiguity, but also redundancy. Specifically, using an explicit DC decreases the potential for ambiguity. For example, the CONTRAST sense in Example 1a is difficult to infer if the DC *but* is omitted. Nonetheless, if the intended discourse sense is highly predictable, it could be verbose or redundant to insert an explicit DC in the utterance, such as the DC *therefore* in Example 1b. In the PDTB, there are similar numbers of implicit and explicit relations (Prasad et al., 2008), yet the corresponding sense distributions are largely different. For example, CONTRAST relations are more common in explicit relations than in implicit relations. These statistics suggest that both options are similarly frequent but the preference is not equally distributed across senses.

In order to explain how human speakers choose the optimal level of marking in their utterances, this work models how speakers rationally balance ambiguity and redundancy.[3] We combine two information-theoretic frameworks, namely the Rational Speech Acts (RSA) model and the Uniform Information Density (UID) principle.

1. In this article, *speakers* and *listeners* are interchangeably used with *authors* and *readers*, respectively.
2. In this work, we use the term "*explicit relations*" to refer to discourse relations that are signaled by explicit DCs.
3. However, this work does not explore the level of consciousness during the reasoning of the marking strategy. We leave it for future work to assess whether people intentionally or subconsciously balance ambiguity and redundancy.

On the one hand, the RSA model (Frank and Goodman, 2012) formalizes the inter-relation of language comprehension and production in terms of a listener model and speaker model, which are interwoven. Recent findings in human language processing suggest that listeners simulate how an utterance is produced to guide comprehension, and speakers consider the ease of comprehension when planning production (Clark, 1999; Kilner et al., 2007; Pickering and Garrod, 2007). Based on these findings, the RSA model quantifies the *informativeness* of the choice of discourse marking by the likelihood for the listerners to disambiguate the discourse relation.

On the other hand, the Uniform Information Density (UID) principle (Levy and Jaeger, 2006) is applied to model how redundant utterances are avoided. The UID principle views language communication as a form of information transmission through a noisy channel, through which a constant rate of information flow is optimal according to Shannon's Information Theory (Genzel and Charniak, 2002; Levy and Jaeger, 2006; Shannon, 1948). Speakers thus structure utterances by optimizing *information density*, which is the quantity of information (measured by *surprisal*) transmitted per *unit of utterance*, typically *a word*. In particular, a highly predictable utterance triggers a drop in *information density*, which has to be smoothed by choosing a more ambiguous utterance, such as by leaving out linguistic markers.

In short, our computational model implements Grice's *Maxim of Quantity* by computing how speakers try to be informative (using the RSA model), but not too informative (based on the UID principle). We apply this model to predict whether an explicit or implicit DC is used to express a discourse relation, given the context of the discourse relation and the discourse sense to be conveyed. Using the actual presence or absence of DCs in the PDTB as the gold standard for evaluation, our model not only achieves a higher accuracy than previous work (Patterson and Kehler, 2013), but also provides an interpretable account of the various cognitive factors behind the predicted decision.

In terms of application, a model that predicts the marking of discourse relations not only contributes to a better understanding of the human language production mechanism, but is also important for automatically generating coherent, human-like text and dialogue. In particular, the degree of marking in discourse relations is cross-linguistically different (Meyer and Webber, 2013; Yung et al., 2015). It remains a challenge for machine translation systems to explicitate (translate an implicit DC to an explicit DC) or implicitate (translate an explicit DC to an implicit DC) discourse relations in source texts, as human translators do (Hoek et al., 2015; Hoek and Zufferey, 2015; Li et al., 2014; Meyer and Webber, 2013; Yung et al., 2015), as it is not yet clear how DC explicitation and implicitation are subject to the convention of discourse marking in the target text.

The rest of this paper is organized as follows. We start with a review of related work on discourse relation marking in Section 2, followed by a description of our proposed methodology in Section 3. Experiments and evaluation using the PDTB are described in Section 4. Lastly, Section 5 discusses the advantages and disadvantages of the methodologies in this study and directions for future work, and Section 6 draws the paper's conclusion.

2. Related work

This section summarizes previous work on modeling the explicit marking of discourse relations. We first give a brief summary on the annotation strategy of the PDTB, which is used as the gold standard for discourse marking in our experiment. We then introduce the state-of-the-art method for the automatic prediction of the presence of discourse connectives in a corpus. Lastly, we describe how discourse relation marking is explained by the UID principle in the existing literature.

On the other hand, background information on the RSA model is explained in Section 3, in connection with our proposed methodology.

2.1 Penn Discourse Treebank (PDTB)

In this work, we applied a computational model to predict the actual marking of discourse relations in corpus data given a particular discourse relation. To achieve this, a corpus annotated with discourse relations and marking is essential. There are various corpora annotated with discourse relations, such as the RST Discourse Treebank (Carlson et al., 2002) and Discourse GraphBank (Wolf et al., 2005), but discourse markers are annotated and associated with discourse relations in only two resources: the PDTB (Prasad et al., 2008) and the RST Signaling Corpus (Das et al., 2015). The proposed model in this work is trained and evaluated against the annotation of the PDTB, which is the largest available discourse-annotated corpus in English.

The PDTB consists of news articles collected from the *Wall Street Journal*. Discourse relations are annotated between each pair of arguments, which are mostly clauses or single sentences. Below are three examples of this annotation.

(2) The OTC market has only a handful of takeover-related stocks. **But** (Explicit; COMPARISON.CONTRAST) they fell sharply. *(WSJ2379)*

(3) Japan's Finance Ministry had set up mechanisms to limit how far futures prices could fall in a single session and ... to give market operators the authority to suspend trading in futures at any time. (Implicit: **but**; COMPARISON) Maybe it wasn't enough. *(WSJ0097)*

(4) This cannot be solved by provoking a further downturn; reducing the supply of goods does not solve inflation. (Implicit 1: **instead** EXPANSION.ALTERNATIVE.CHOSEN ALTERNATIVE), (Implicit 2: **so**; CONTINGENCY.CAUSE.RESULT and EXPANSION.ALTERNATIVE) Our advice is this: Immediately return the government surpluses to the economy...

The PDTB follows a lexically-grounded approach in the annotation of discourse relations (Webber et al., 2003). First, *explicit* DCs are identified, based on a list of DCs that are accumulated in the course of annotation, and labeled with relation senses (Example 2). Other expressions that signal discourse relations, such as "the reason is", are identified as *alternative lexicalization (AltLex)* and labeled with relation senses as well. If explicit markers are absent *between two sentences* within the same paragraph, there are three options for annotation: i) if a discourse relation can be inferred and expressed by a DC, the relation is labeled as *implicit* and the candidate DC and relation sense are annotated (Example 3); ii) if a discourse relation cannot be inferred but the two sentences are about the same entity, the relation is labeled *EntRel*; and iii) if the two sentences are unrelated, they are tagged as *NoRel*. This work focuses on the marking of discourse relations by discourse connectives, so we only make use of samples labeled *explicit* or *implicit*.

Senses in the PDTB are defined in a hierarchy of two to three levels, as shown in Figure 1. Some relations have multiple senses. Up to two DCs can be assigned to an implicit relation and, in turn, each (implicit or explicit) DC can be labeled with up to two senses (Example 4). Similarly, certain level 2 senses, as in Example 2, are resulting from the back-off strategy in annotation, i.e. when the annotators disagree on the level 3 senses. This is also a kind of multiple sense. We argue that multi-sense discourse relations are non-compositional, which means that we consider each combination of multiple senses as an individual sense. More details are given in Section 4.1.

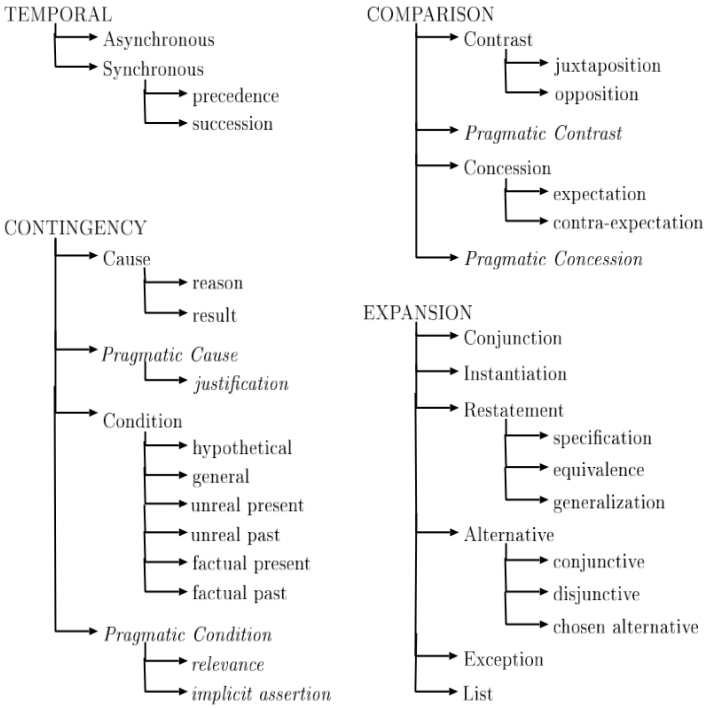

Figure 1: Sense hierarchy of PDTB (Prasad et al., 2008)

2.2 Automatic prediction of discourse relation marking

The choice of discourse marking strategies has been studied in earlier works as a subtask for natural language generation (Allbritton and Moore, 1999; Danlos, 1998; Grote and Stede, 1998; Moser and Moore, 1995; Scott and de Souza, 1990; Soria and Ferrari, 1998). In the absence of large-scale resources, investigations are based on manually derived rules and lexicons or psycholinguistic experiments.

With the emergence of large corpora annotated with discourse relations, Patterson and Kehler (2013) presented a machine-learning approach to predict whether an explicit or implicit DC is used in the corpus for a particular discourse relation. They argue that while the choice is related to the ease of inference, it may also depend on other stylistic or textual factors. A classifier is trained to predict whether a *candidate DC* (the DC that actually occurs in the text as an explicit DC or annotated as an implicit DC) is actually present, given the sense of the discourse relation and the arguments. Features include observable surface forms, such as argument length, count of subject nouns, and content word ratio, as well as contextual discourse structures, such as the previous discourse relation and whether the relation is embedded or shared. The classifier is trained and tested on a subset of the most frequent relations from the PDTB. An overall high classification accuracy is achieved and relation-level and discourse-level features are found to be more useful than argument-level features.

We evaluate our proposed discourse marking model by predicting the use of explicit or implicit DCs in PDTB, as in Patterson and Kehler (2013). However, Patterson and Kehler (2013) focus

on a data-driven approach that correctly replicates the occurrence of DCs in the corpus without a theoretically grounded explanation of why an utterance is preferred by the speaker. Our work differs in that we model the option of marking from the viewpoint of human language production, explaining the speaker's choice in terms of information theories. As a model of human language production, we do not make use of the *candidate DC* to represent the message to be conveyed by the speaker, as it is the result of the speaker's choice, if an explicit DC is preferred. We only make use of the relation sense label to represent the message. Nonetheless, our model achieves higher accuracy when evaluated on the same test samples.

2.3 Discourse relation marking explained by UID

The UID principle provides a theoretical basis that connects the use of DCs with that of other discourse relation signals. According to UID, information density rises when an utterance is "surprising" and drops when an utterance is highly predictable. To smooth the peaks and troughs, speakers adjust the ambiguity of an utterance by including or omitting linguistic markers.

UID is applied to explain a variety of speaker's options, such as phonetic (Aylett and Turk, 2004), morphological (Frank and Jaeger, 2008), and syntactic (Jaeger, 2010) reductions as well as referential expressions (Tily and Piantadosi, 2009). In the context of discourse relations, our approach incorporates the UID assertion that explicit DCs are omitted when the discourse relation is highly predictable.

An analysis of the PDTB in the literature shows that CAUSAL and CONTINUOUS senses are more often implicit, or marked by less specific DCs (Asr and Demberg, 2012). Indeed, these senses are presupposed by listeners according to linguistics theories (Kuperberg et al., 2011; Levinson, 2000; Murray, 1997; Sanders, 2005; Segal et al., 1991). In addition, the DC *instead* is more often dropped for the discourse relation CHOSEN ALTERNATIVE, if the first argument contains negation words, which are identified cues for this relation (Asr and Demberg, 2015). In fact, expectations about discourse relations are triggered by various signals, such as verb classes (Rohde and Horton, 2014).

The corpus statistics presented in these analyses support the UID hypothesis that expected, predictable relations are more likely to be conveyed implicitly, and thus more ambiguously, to maintain a steady information flow. However, there are explicit CAUSAL and CONTINUOUS relations and some CHOSEN ALTERNATIVE are marked even the first argument is negated. Although measures have been proposed to rate the implicitness of a relation sense (Asr and Demberg, 2013; Jin and de Marneffe, 2015), these measures only quantify the general marking of each sense in the data (e.g., the CONTRAST sense), but not the speaker's choice for each particular instance (e.g., the CONTRAST sense, given particular arguments and context).

In contrast, our model incorporates an *information density predictor*, which specifically predicts the expectability of a given relation. In turn, the speaker's choice of discourse marking is biased based on the predicted degree of expectability. Instead of particular senses or cues in the corpus, we generally apply UID to model each relation instance of in the corpus irrespective of the relation sense, in conjunction with other language production factors.

Our research questions are as follows:

1. Does our proposed model explain speakers' choice of DC marking? If the hypotheses of the model is appropriate, each component in the model should contribute to the prediction accuracy.

2. How does the prediction performance of the proposed model compare with the state-of-the-art, i.e., Patterson and Kehler (2013)?

3. Modeling the marking of discourse relations using the RSA model

This section describes our proposed method for modeling the speaker's choice of DC marking. We start by explaining the key elements of the RSA model. Then, in order to provide a full picture of the application of RSA to discourse processing, we also give a brief account of discourse relation interpretation using the *listener model* of RSA, as described in Yung et al. (2016). This is followed by the details of our proposed marking model, which predicts the marking of a discourse relation produced by a speaker, and is based on the *speaker model* of RSA.

3.1 Theoretical framework: the Rational Speech Acts (RSA) model

The RSA model (Frank and Goodman, 2012) describes the speaker and listener as rational agents who cooperate towards efficient communication in a game-theoretic approach. The model belongs to the line of reasoning that speakers and listeners cooperate in a conversation by recursively inferring the reasoning of each other (Benz et al., 2016; Frank and Goodman, 2012; Goodman and Lassiter, 2014; Goodman and Stuhlmüller, 2013; Jäger, 2012), and studies of Bayesian interpretation in general (Zeevat, 2011, 2015).

RSA has been used to successfully explain existing psycholinguistic theories and predict experimental results at various linguistic levels, such as the perception of scalar implicatures (e.g., "some" meaning "not all" in pragmatic usage) and the production of referential expressions (e.g., using pronouns or proper nouns to refer to an entity) (Bergen et al., 2014; Kao et al., 2014; Lassiter and Goodman, 2013, 2015; Potts et al., 2015). Recent efforts also learn and evaluate the models using corpus data instead of experimental data (Monroe and Potts, 2015; Orita et al., 2015).

Based on the theory that production guides comprehension and that comprehension guides production, the RSA model is composed of a listener model embedding a speaker model and a speaker model embedding a listener model. Details are explained in the following sections.

3.1.1 LISTENER MODEL

Rational listeners assume the utterance they hear contains the optimal amount of information. Listeners predict the intended message of speakers by Bayesian inference (Equation 1),

$$P_{listener}(s|w, C) \propto P_{speaker}(w|s, C)P(s) \tag{1}$$

where s is the *meaning* of an utterance; w is the *utterance* produced by the speaker, and C is the *context*.

During comprehension, the listener reasons about how the utterance is produced. $P_{speaker}(w|s, C)$ represents the speaker model *in the listener's mind*. On the other hand, $P(s)$ represents the *salience* of the meaning, which is the private preference of the listener, but is also subject to the shared knowledge between the speaker and listener.

3.1.2 SPEAKER MODEL

Rational speakers emulate the listeners' interpretation and choose an utterance they believe to be informative. In addition, an utterance that is easy to produce is preferred. Specifically, speakers choose an utterance by soft-max optimizing the expected *utility* ($U(w; s, C)$) of the utterance (Equation 2).

$$P_{speaker}(w|s, C) \propto e^{\alpha \cdot U(w;s,C)} \tag{2}$$

Utility is the effectiveness of the speaker using utterance w to express meaning s in context C, and α is the decision noise parameter, which is set to 1 to represent a rational speaker. Here, $\alpha = 0$ means the decision is completely unrelated to pragmatic reasoning; $\alpha = 1$ represents the Luce's choice axiom (Frank and Goodman, 2012), i.e., a rational decision; and $\alpha > 1$ suggests biased choices.

The speakers select an utterance which, they think, is informative to the listeners and not costly to produce. *Utility* is thus defined as the *informativeness* ($I(s; w, C)$) of the utterance, deducted by the cost ($D(w)$) to produce it (Equation 3).

$$U(w; s, C) = I(s; w, C) - D(w) \tag{3}$$

If the sense to be conveyed by the chosen utterance is unconventional and surprising, the utterance is less useful. Therefore, *informativeness* is quantified by the *negative surprisal* of the sense to be conveyed with respect to the utterance being used (Equation 4).

$$I(s; w, C) = \ln P_{listener}(s|w, C) \tag{4}$$

In turn, during discourse production, the speaker emulates how the listener interprets the utterance. Here, $P_{listener}(s|d, C)$ is the listener model *inferred by the speaker*. It is the probability that the *speakers assume* the listeners can interpret their intended meaning s.

To summarize, the speaker and listener emulate the language processing of each other. However, instead of unlimited iterations (i.e., the speaker thinks the listener thinks the speaker thinks...), the inference is grounded by a literal interpretation of the utterance. Literal interpretation means the listener does not reason about the the likelihood of the sense of an utterance, but always assigns the same interpretation to the same utterance. Similarly, a literal speaker always uses the same utterance for the same sense.

Figure 2 illustrates the direction of pragmatic inference between the speaker and listener *in their minds*. Literal listener L_0 and literal speaker S_0 do not reason about the reasoning of their counterparts. Pragmatic listener L_1 reasons about the pragmatic speaker S_1, who in turn reasons about literal listener L_0. Pragmatic listeners and speakers at higher levels (e.g., L_2 and S_2) reason with more iterations, but previous studies demonstrate that one level of reasoning is robust for modeling a human's interpretation (Goodman and Stuhlmüller, 2013; Lassiter and Goodman, 2013; Yung et al., 2016). Our proposed model for DC production is based on pragmatic speaker S_1, who reasons about literal listener L_0.

3.2 Listener model for discourse relation interpretation

Yung et al. (2016) used the L_1 listener model of RSA to model how listeners interpret the sense of a DC. Given DC w and context C in a text, the listener's interpreted relation sense s_i is the sense that

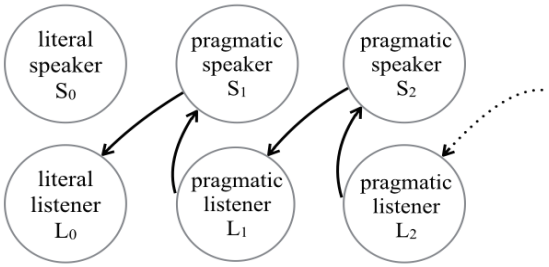

Figure 2: Directions of reasoning of listeners and speakers. (Reproduced from Yung et al. (2016))

maximizes $P_{listener}(s|w, C)$,and s_i is specifically defined as

$$s_i = \arg\max_{s \in S} P_{listener}(s|w, C) \tag{5}$$

where S is the set of defined relation senses. Context variable C is defined by the immediately previous relation, including the sense and form (explicit DC or not) of the relation.

First, literal listener L_0 interprets the DC directly by its most likely sense in the context. The probability is estimated by counting the co-occurrences in the PDTB in which explicit and implicit DCs are labeled with discourse relation senses.

$$P_{L_0}(s|w, C) = \frac{count(s, w, C)}{count(w, C)} \tag{6}$$

Next, pragmatic speaker S_1 estimates the utility of a DC by emulating the comprehension of the literal listener L_0 (Equations 2, 3, and 4). The probability that pragmatic speaker S_n will use DC w to express meaning s is estimated as:

$$P_{S_n}(w|s, C) = \frac{e^{\ln P_{L_{n-1}}(s|w,C) - D(w)}}{\sum_{w' \in W} e^{\ln P_{L_{n-1}}(s|w',C) - D(w')}} \tag{7}$$

where $n \geq 1$. Moreover, W is the set of annotated DCs, including *null*, which stands for an implicit DC.

The cost function $D(w)$ in Equation 7 measures the production effort of the DC. Yung et al. (2016) simply defined the cost of producing *any explicit DC* by a constant positive value, which is tuned manually in the experiments, while the production cost for an implicit DC is 0, since no word is produced. In this work, we further develop other measures to quantify the effort of the utterance. This is explained in Section 3.3.4.

Finally, pragmatic listener L_1 emulates the DC production of pragmatic speaker S_1 (Equation 1). The probability that the pragmatic listener L_n will assign meaning s to DC w is estimated as:

$$P_{L_n}(s|w, C) = \frac{P_{S_n}(w|s, C)P_L(s)}{\sum_{s' \in S} P_{S_n}(w|s', C)P_L(s')} \tag{8}$$

where $n \geq 1$ and S is the set of defined senses. The salience of a relation sense $P_L(s)$ is defined by the frequency of the sense in the corpus.

$$P_L(s) = \frac{count(s)}{\sum_{s' \in S} count(s')} \tag{9}$$

The RSA model argues that rational listeners do not just stick to the literal meaning of an utterance. Instead, they reason about how likely it is that the speaker will use that utterance, in the current context, based on the informativeness and production effort of the utterance. To evaluate this claim, Yung et al. (2016) compared the DC interpretation by the literal listener (based on the probability estimates of P_{L_0}, Equation 6), and the pragmatic listeners (based on the probability estimates of P_{L_1} and P_{L_2}, Equation 8). Their experiments based on the PDTB find that discourse sense predictions made by the pragmatic listeners outperform predictions by the literal listener, providing support to the RSA model.

3.3 Proposed method to model discourse relation production

In contrast to Yung et al. (2016), who applied the *listener model* of RSA to model the human comprehension of discourse connectives, this work applies the *speaker model* of pragmatic speaker S_1 to model the production of discourse structure. Specifically, the proposed model predicts whether the speaker will use an explicit or implicit DC given the discourse relation to be conveyed. We first explain how we adapt the RSA model to predict discourse relation marking, followed by the details of each component.

3.3.1 SPEAKER MODEL AS A MARKING MODEL FOR DISCOURSE RELATIONS

The probability pragmatic speaker S_1 will use utterance w to convey intended message s in context C is:

$$P_{S_1}(w|s,C) = \frac{e^{U(w;s,C)}}{\sum\limits_{w' \in W} e^{U(w';s,C)}} \tag{10}$$

We consider the binary choice of explicit or implicit DCs in this task. Utterance w thus comes from set W = {(exp)licit, (imp)licit}, where both explicit and implicit DCs are **grammatically valid** to convey s, the sense of discourse relation. Our model thus predicts a speaker's choice of DCs based on the following two probabilities:

$$\begin{aligned} P_{S_1}(exp|s,C) &= \frac{e^{U(exp;s,C)}}{e^{U(exp;s,C)} + e^{U(imp;s,C)}} \\ P_{S_1}(imp|s,C) &= \frac{e^{U(imp;s,C)}}{e^{U(exp;s,C)} + e^{U(imp;s,C)}} \end{aligned} \tag{11}$$

According to Equation 3, the *utility* U of an explicit DC equals its *informativeness* I less the production cost D. We define the *informativeness* of using an explicit DC as the difference in the amount of information conveyed when the DC is used and not, which is quantified by negative *surprisal* [4].

$$\begin{aligned} U(exp;s,C) &= I(s;exp,C) - D(exp) \\ I(s;exp,C) &= \ln P_{L_0}(s|exp,C) - \ln P_L(s|C) \end{aligned} \tag{12}$$

where $P_L(s|C)$ is the *salience* of sense s in context C, irrespective of how the sense is presented. High $I(s;exp,C)$ means it is informative and not surprising to use an explicit DC for this sense.

4. The conventional term *informativeness* in RSA is defined by *negative surprisal*, while *information density* in UID is the *surprisal*.

In contrast, if the speaker does not use an explicit DC, the relation sense is inferred from the arguments. In addition, as discussed in Section 2.3, default discourse senses are more often unmarked. In other words, a *null* DC is also informative for discourse sense prediction.

We thus define the probability that a speaker will choose an implicit DC to be proportional to the *sum of the the utilities* of a *null* DC and arguments ($args$). Therefore:

$$\begin{aligned} e^{U(imp;s,C)} &= e^{U(null;s,C)} + e^{U(args;s,C)} \\ U(null;s,C) &= I(s;null,C) - D(null) \\ U(args;s,C) &= I(s;arg,C) - D(args) \end{aligned} \tag{13}$$

No effort is required to produce a *null* DC. Further, we assume that the arguments have been produced to convey other information irrespective of their discourse informativeness, so no extra effort is needed. Therefore, $D(null)$ and $D(args)$ both equal 0.

The amount of information that the null DC provides for the discourse relation is defined similarly to Equation 12, as follows:

$$I(s;null,C) = \ln P_{L_0}(s|null,C) - \ln P_L(s|C) \tag{14}$$

The informativeness of the arguments is used as the *information density predictor*, following the UID principle. It is less straightforward to measure. We propose an indirect measure that we explain in detail in Section 3.3.3.

To summarize, the marking model predicts that speakers will use an explicit DC if

$$e^{U(exp;s,C)} > e^{U(null;s,C)} + e^{U(args;s,C)} \tag{15}$$

and that they will use an implicit DC otherwise. Details of each of the components are explained in the following sections.

3.3.2 INFORMATIVENESS OF DCS

This section explains how we estimate the informativeness of using an explicit and implicit DC respectively in Equations 12 and 14. We can assume that the utterance lexicon $W = \{exp, imp\}$ in Equation 10 and the set of speaker's intended messages (all possible discourse relation senses) are always *valid*. This is in contrast with other pragmatic situations where RSA has been applied. For referential expressions, for example, the lists of referents and grammatically correct pronouns differ case by case.

As in the interpretation model in Section 3.2, we extract the universal distributions of $P_L(s|C)$, $P_{L_0}(s|exp,C)$, and $P_{L_0}(s|null,C)$ from corpus data. This is based on counting co-occurrences between a sense and the occurrence of an explicit or null marker under a context. We extract these distributions from the training portion of the corpus, i.e., excluding the samples for testing and tuning parameters (see Section 3.3.3).

Following Yung et al. (2016), we define context C as the surrounding discourse relations, which are also used as features in discourse planning for natural language generation (Biran and McKeown, 2015). Specifically, the discourse contexts are: the full discourse sense annotated in PDTB (S), the fourway top level sense (TS), the form of discourse presentation (F) such as "explicit" or "implicit", the combination of sense and form (SF), and the combination of top sense and form (TSF). [5] The

5. We use the five forms of discourse presentation defined in the PDTB: explicit DC, implicit DC, alternative lexicalization, entity relation and "no relation".

contexts are taken from window sizes of 1 to 2: previous one (10), next one (01), previous two (20), next two (02), and the previous one paired with the next one (11). We hypothesize that the speaker also thinks ahead about the coming discourse structures when planning the current ones. Predictions based on various discourse contexts are compared in the experiment.

3.3.3 INFORMATIVENESS OF ARGUMENTS

The informativeness of arguments $I(s; arg, C)$ in Equation 13 refers to the contribution of the arguments to present the discourse sense. It estimates the information density of the utterance towards the sense. Following the UID principle, information density drops when the intended discourse sense is predictable from the arguments alone, and thus the explicit DC is omitted.

The presence of features in the arguments that signal a particular sense makes the sense more predictable, and thus reduces the chance that an explicit DC for that sense is used. For example, as introduced in Section 2.3, the DC *instead* is less used to present the CHOSEN ALTERNATIVE sense if the first argument is negated (Asr and Demberg, 2015).

In order to model the marking of every relation, we generalize the idea to capture various cues in the arguments for all senses by means of an automatic discourse parser. The implicit DC sense classifier of the discourse parser uses various features in the arguments to identify implicit discourse senses (Lin et al., 2009; Park and Cardi, 2012; Pitler et al., 2009; Rutherford and Xue, 2014). For example, if there is a pair of antonyms in the two arguments (e.g. "high" in the first argument, and "low" in the second argument), the CONTRAST relation is more likely. Furthermore, modal words, such as *should* and *may*, can be used to express the CONDITIONAL relation.

Given a pair of arguments, the classifier estimates the probability of each discourse sense and outputs the most likely sense. A high probability estimate indicates a more certain sense prediction, suggesting that more cues are identifiable from the arguments, and thus explicit marking is less necessary. Our motivation for using the implicit DC classifier is based on the hypothesis that the classifier can better predict the sense of relations that are actually implicit than those that are actually explicit because more features in the arguments are identifiable.

Similar to the informativeness of DCs, we quantify the informativeness of the arguments using an information-theoretic approach. For comparison, we propose two methods to approximate the informativeness of the arguments based on the probability distribution estimated by an automatic implicit DC parser: (1) the *negative surprisal* of the estimated probability P_p of the parser output sense s_{output} (Equation 16), and (2) the *negative entropy* of the probability distribution estimated by the parser (Equation 17).

$$I(s; arg, C) = w_a \cdot \ln P_p(s_{output}) \tag{16}$$

$$I(s; arg, C) = w_a \sum_{s_p \in O} P_p(s_p) \log P_p(s_p) \tag{17}$$

where O is the set of senses defined in the parser and w_a is a positive weight tuned on the development samples (a set of held-out samples different from the training and testing sets). We measure the *general informativeness* of the arguments to imply *any* discourse senses, so s_{output} does not necessarily equal s.

We employ the implicit sense classifier from the winning parser of the CoNLL shared task 2015 (Wang and Lan, 2015), which was designed to identify a subset of fourteen implicit senses plus the *entity relation*. Features used in the classifier include production rules, dependency rules, last word

or argument 1, first three words of argument 2, presence of modality verbs and inquirer, polarity, the immediately preceding DC, and Brown cluster pairs. In our experiment, the syntactic features are based on automatic parsing using the Stanford CoreNLP (Manning et al., 2014).

The two arguments of a relation instance, which can actually be explicit or implicit, are passed to the implicit DC classifier and $I(s; arg, C)$ is calculated using the output probabilities. The parser has been trained on the same sections of the PDTB as the training set used in our experiment, so there is no overlap with our test samples. Although the performance of this state-of-the-art implicit DC classifier is still unsatisfactory (34.45% on PDTB Section 23)[6], our method only makes use of the confidence of the prediction, which is based on whether discourse related features are detected or not.

We hypothesized that the implicit DC classifier can better predict actually implicit relations than actually explicit relations. In fact, this is the case. The classification accuracy of the originally explicit relations (28.45%) is significantly lower than that of the originally implicit relations (51.30%) on the test set, when matching at the fourway top level discourse sense and counting predictions of *entity relation* as *Expansion*. This supports our motivation to use the parser estimation as an information density predictor.

3.3.4 COST FUNCTION

Cost function $D(exp)$ models a speaker's effort required to produce an explicit DC for the intended discourse sense. We propose five versions of the cost function that are inspired by existing psycholinguistic findings.

1. **Mean DC length**
   Production cost intuitively increases with word length (Orita et al., 2015). We define the mean DC length of a discourse relation as the mean number of characters of all valid DCs for that sense normalized by the average word length of all DCs. A lexicon of possible DCs per discourse sense is derived from the whole corpus. For multi-word DCs, a white space is simply counted as one character. As mentioned in Section 2.3, we view that speakers first decide to use an explicit DC or not, then decide which DC best expresses the relation. Therefore, we do not use the length of the *candidate DC* directly.

2. **DC/arg2 ratio**
   Similarly, we use the mean number of words normalized by the number of words in *argument 2* as another version of the cost function. This is based on the hypothesis that DC production cost is related to the relative length of the DC comparing with the whole utterance, rather than the absolute length.

3. **Prime frequency**
   Structural priming refers to the tendency for humans to process a linguistic construction (the target) more easily if the construction has been used before. In terms of language production, a speaker tends to repeat a previous construction (the prime) because it takes less effort than generating an alternative construction.
   A lexical prime of an explicit DC is ideally the same explicit DC or an explicit DC of the same sense. However, we found that repetitions of the same DC are too sparse in the data.

6. http://www.cs.brandeis.edu/~clp/conll15st/results.html

Therefore, we define the choice of explicit or implicit marking as the structural prime of presenting a discourse relation. Specifically, we use the reciprocal of the count of primes (any explicit DC occurring before the current position) as the production cost, since the strength of priming effect is known to increase with the frequency of the primes (Bock, 1986; Levelt and Kelter, 1982; Smith and Wheeldon, 2001).
It is not yet clear whether priming occurs in DC selection. For example, it is also reported that speakers tend to avoid the same DC if the relation is embedded in a relation preceded by the same DC (Moser and Moore, 1995). The result of our experiment could provide indirect evidence on discourse level priming.

4. **Prime distance**
We also use the prime-target distance normalized by the length of the text article as another version of the production cost. Psycholinguistic findings suggest that the priming effect is more subtly affected by the prime-target distance (Bock et al., 2007; Gries, 2005; Jaeger and Snider, 2008).

5. **Distance from start**
We use the relative position of the relation within the text as the production cost. We hypothesize that more effort is needed as the production proceeds, as the accumulated length of the whole utterance increases.

The range of values of the cost function depends on the cost definition. We thus adjust the values with a constant weight w_c, which is tuned on the development samples in the experiments:

$$D(exp) = w_c \cdot cost(exp) \tag{18}$$

4. Experiment

We apply the model to simulate a speaker's choice of explicit or implicit DC for discourse relations in the PDTB corpus. The aim of the experiment is to answer our research questions introduced in the end of Section 2.

4.1 Data and setting

The experiment is based on the annotation of discourse relation senses and explicit/implicit DCs in the PDTB, as described in Section 2.1. Annotations of other forms of discourse relations, such as entity relations and attributions, are excluded. In addition, our model is based on the assumption that $W = \{explicit, implicit\}$ for all relations, yet it is notable that *intra-sentential implicit* DCs are **not** annotated in the PDTB (Prasad et al., 2014). In addition, as a result of the annotation procedure, implicit relations always occur *in between two arguments*. We thus exclude intra-sentential samples and cases where the DCs are not in between two arguments, hence $W = \{explicit, implicit\}$ is always true. The resulting experimental data set contains $5,201$ explicit and $16,049$ implicit relations[7].

Most existing works split a multi-sense sample into separate samples, each labeled with one of the senses. However, it is notable that the individual senses of a multi-sense relation are not disjoint

7. Four cases of intra-sentential implicit relations, due to sentence splitting errors in the corpus, are removed. In the testing phrase, excluded samples are counted as *explicit* by default.

and *having multiple senses* is *part of the sense* (Asr and Demberg, 2013; Prasad et al., 2014). The property of multiple senses is an important factor of our DC production model: speakers could choose an explicit DC for each sense, but if they have to express two senses at the same time, an implicit DC could be more usable. Therefore, we treat all combinations of senses as *individual senses*, each containing one to three joint sense labels.[8] This results in a total of 122 senses.

Table 1 is a summary of the distribution of the relation senses in descending order of frequency. In fact, joint multi-senses are not rare: the most frequent multi-sense, EXPANSION.CONJUNCTION–TEMPORAL.SYNCHRONY, is the 17th most frequent sense.

	Sense	Exp	Imp
1	Expansion.Conjunction	1,380	3,314
2	Comparison.Contrast	1,283	1,200
3	Expansion.Restatement.Specification	75	2,406
4	Contingency.Cause.Reason	28	2,295
5	Contingency.Cause.Result	269	1,649
6	Expansion.Instantiation	119	1,383
7	Comparison.Contrast.Juxtaposition	507	672
8	Comparison.Concession.Contra-expectation	475	179
9	Temporal.Asynchronous.Precedence	117	479
10	Expansion.List	84	374
...	...	...	...
17	Expansion.Conjunction#Temporal.Synchrony	74	114
...	...	...	...
20	Expansion	8	89
...	...	...	...
50	Contingency.Pragmatic cause.Justification#Expansion.Instantiation	0	6
...	...	...	...
122	Contingency	0	1
Total		**5,201**	**16,049**

Table 1: Sense distribution of explicit and implicit DCs in the experimental data.

The experimental data are split in the same way as in previous work (Patterson and Kehler, 2013): sections 2–22 are used as the training set, sections 0–1 as the development set; and sections 23–24 as the test set. In the training phrase of the experiment, probability distributions in the marking model are deduced from the training set of the experimental data. In the testing phrase, the model is applied to predict whether an explicit/implicit DC is likely to be used for each discourse relation in the development set and test set. For direct comparison with previous work, samples of infrequent DCs and relation senses were excluded from the development and test sets according to the same criteria as in previous work (Patterson and Kehler, 2013). The resulting development and test sets contain $1,720$ and $1,878$ relations, respectively. During evaluation, the predictions are compared with the actual marking in the corpus. Parameters in the model (w_a and w_c) were selected

8. There is only one sample of three joint labels in our experimental dataset (Example 4), although up to four joint labels are possible (two implicit DCs labeled with two senses each).

to maximize the prediction accuracy on the development set and the same optimal parameters were used on the test set.

4.2 Results

The explicit and implicit DC prediction performance of our proposed marking model based on various component combinations of the model is summarized in Table 2.

	Discourse context C	Arg. info. $e^{U(args;s,C)}$	Cost function $D(exp)$	Dev: Sections 0-1 Accuracy	Dev: Sections 0-1 $F1_{exp}$	Dev: Sections 0-1 $F1_{imp}$	Test: Sections 23-24 Accuracy	Test: Sections 23-24 $F1_{exp}$	Test: Sections 23-24 $F1_{imp}$
BL	constant	0	0	.8494	.8721	.8170	.8536	.8754	.8225
SOA (Patterson and Kehler, 2013)				–	–	–	.8660	–	–
(a)	F10	0	0	.8552	.8762	.8258	.8552	.8760	.8259
	SF10	0	0	.8593	.8773	.8351	.8546	.8736	.8291
	F20	0	0	.8541	.8748	.8251	.8541	.8748	.8253
	F11	0	0	.8512	.8723	.8217	.8541	.8749	.8250
	TS10	0	0	.8523	.8723	.8217	.8536	.8748	.8236
(b)	constant	surprisal	0	.8948^{++}	.9013	.8874	.8701	.8810	.8570
	constant	entropy	0	.8953^{++}	.9019	.8879	.8695	.8805	.8563
(c)	constant	0	mean DC length	.8936^{++}	.8968	.8902	.8759^{+}	.8855	.8646
	constant	0	DC/arg2 ratio	.8948^{++}	.9004	.8885	.8733	.8821	.8631
	constant	0	prime frequency	.8860^{+}	.8879	.8851	.8727	.8821	.8618
	constant	0	prime distance	.8924^{++}	.9018	.8812	.8749	.8855	.8620
	constant	0	distance from start	.8930^{++}	.8938	.8923	.8770^{+}	.8790	.8749
(d)	F10	entropy	DC/arg2 ratio	**.9017**$^{++}$	**.9025**	**.9010**	.8818^{+}	.8829	.8806
	TSF01	surprisal	prime frequency	.8953^{++}	.8983	.8922	.8892^{++*}	**.8931**	.8851
	TS01	entropy	prime distance	.8948^{++}	.8998	.8892	**.8903**$^{++*}$	.8923	**.8883**

Table 2: Accuracies and F1 scores of predicted DC marking. The best values are bolded. (abbreviations: S: full relation sense; TS: top-level sense; F: relation form; SF: sense and form; TSF: top sense and form; 10: previous relation; 20: previous 2 relations; 11: previous relation and next relation)
$^{+/++}$: significant improvement over baseline (**BL**) accuracy at $p < .05$ and $p < .001$ respectively;
*: significant improvement over state-of-the-art (**SOA**) accuracy at $p < .03$ (by Pearson's χ^2 test)

Row BL shows the results of the model without the cost function and argument informativeness component with constant context C. We considered this setting to be the baseline, in which the prediction is solely based on the distributions of $P(s|exp)$ and $P(s|imp)$. Considerably high accuracy is achieved, suggesting that the speaker's choice of marking is strongly related to the intended discourse sense.

Row (a) shows the prediction results based on the distributions of $P(s|exp, C)$ and $P(s|imp, C)$, where C is the discourse context. The five best combinations of contexts and window sizes are shown. Refining the utility of DCs using these contextual constraints, such as previous relation

senses and marking, does not improve the classification accuracy. This suggests that a speaker's choice of marking depends on other contextual factors rather than surrounding discourse relations.

Row (b) shows the contribution of the *argument informativeness* component, under a constant discourse context and production cost. Classification accuracy increases (significantly for the development set) when the usability of an explicit DC is reduced by the estimated informativeness of the arguments, supporting the UID principle. Predictions based on the surprisal of the parser output sense and entropy of the parser output distribution are similar. We also experiment by adjusting the estimated *argument informativeness* only if the parser output sense is correct (matching at the top level sense). Similar improvement is observed.

Row (c) shows the contribution of the cost function, when the discourse context is set as a constant and argument informativeness is not considered. Adjusting the utility of explicit DCs by their production cost increases the classification accuracy significantly. Among the various features to model the production cost, "mean DC length" and "distance from start" features give the best results, while "prime frequency" and "prime distance" are the least effective. This suggests that the cost to produce a DC is more subject to the mean DC length and that a priming effect in DC production may be more subtle comparing with informativeness.

Row (d) shows the performance of predictions based on the three best combinations of components. The highest accuracies and F_1 scores are achieved for both explicit and implicit relations.

These results answer the first question of the experiment: the proposed model explains the speaker's choice of DC marking in terms of DC and argument informativeness, as well as production cost, while contextual discourse structure is not a significant constraint on the choice.

The answer to the second question is also positive. Significant improvement above the state-of-the-art (Row **SOA**) is achieved by the two best combinations (89.03% and 88.92% vs. 86.60%).

Lastly, we compared the results with a linear classifier trained on the *features* specified in the model, i.e., the discrete values of the intended sense and various discourse context definitions, and real values of various cost functions and argument informativeness estimates. Note that in the proposed model, the training data are used to derive the $P(s|exp, C)$ and $P(s|null, C)$ distributions only, while the linear classifier learns from the features and DC marking of the training set. We used LIBLINEAR (Fan et al., 2008) to build the classifiers. When extracting the argument informativeness features from the training set, using the automatic discourse parser, we penalize the parser estimates of the *implicit* samples by a constant ratio, since the discourse parser is also trained on these samples. The classifier achieves an accuracy of 88.3% on the test set, which does not significantly outperform previous work. This suggests that the information-theoretic configuration is an advantage of our model.

5. Discussion

In this work, we proposed a computational model to predict discourse marking in human language production. The model is trained and evaluated using manual discourse annotation on corpus data as the gold standard. This section explains the advantages and disadvantages of our methodology and suggests directions for future work.

One advantage of learning the discourse marking model from PDTB is the compatibility of the PDTB's annotation with the RSA framework. Previous applications of RSA focus on the pragmatic use of language, where the *intended message* and *lexicon of an utterance* largely depend on the context. In the task of referring expression generation, the sets of valid referents and referring ex-

pressions differ case by case. For example, *red* is an invalid option for referring to a *blue* ball; *he* is an invalid option for referring to a *woman*; and it is difficult to define a finite set of referents in the corpus. In contrast, discourse relations are generally universal across different contexts. A DC can be used or dropped to represent various discourse senses in various contexts, while referring expressions are limited by the properties of each particular referent. In addition, the PDTB annotation scheme pre-defines the sets of DCs and discourse relation senses. In this way, the listener and speaker models of RSA can be derived statistically by counting the co-occurrence of the DCs, sense labels, and contextual factors in the annotated corpus. However, our proposal to use surrounding discourse relations as context did not improve classification accuracy. Therefore, one direction to improve the proposed model is to make fuller use of the training data to learn a more expressive and general abstraction of the context governing the choice of discourse marking.

However, the pragmatic reasoning approach of RSA has been criticized for being unrealistic, because previous studies find that speakers tend to produce referring expressions that are over-specifying (Baumann et al., 2014; Dale and Reiter, 1995; Engelhardt et al., 2006; Gatt et al., 2013). In other words, while ideal pragmatic speakers should only focus on the minimum properties that help listeners to identify the referent, the referring expressions that speakers actually choose often include redundant properties that are not necessary for distinguishing the referent from other candidates. In the context of discourse marking, an utterance is over-specifying if an explicit DC is chosen even though an implicit DC is enough for the listeners to infer the discourse relation.

Another advantage of the proposed model is that it does not rely on pragmatic reasoning alone to model speakers' choice of discourse marking, but also makes use of the UID principle to penalize the choice of explicit DCs when informative signals are present in the arguments. In addition, learning RSA from corpus statistics allows the model to detect the general trend in the marking of a relation sense. Some relation senses are highly likely to be marked/unmarked irrespective of the presence of other signals, while for other relations, the presence of other discourse signals affects the choice, as illustrated in Examples 5 to 7. In these examples, the speaker probabilities (Equation 11) estimated by the best performing model (last row in Table 2) are shown along with the predicted marking choices. For comparison, we also show $P'_s(imp|s, C)$ and $P'_s(exp|s, C)$, which are the speaker probabilities without the UID bias (i.e., $e^{U(arg;s,C)} = 0$, in Equation 13).

(5) And market expectations clearly have been raised by the capital gains victory in the House last month (Implicit:**since**; CONTINGENCY.CAUSE.REASON) An hour before Friday's plunge, that provision was stripped from the tax bill. *(WSJ2429)*

(without UID)	$P'_s(exp\|s, C) = 0.042$	$P'_s(imp\|s, C) = 0.958$	
(with UID)	$P_s(exp\|s, C) = 0.024$	$P_s(imp\|s, C) = 0.976$	Prediction= Implicit

(6) Boeing's offer represents the best overall three-year contract of any major U.S. industrial firm in recent history. **But** (Explicit; COMPARISON.CONTRAST.OPPOSITION) Mr. Baker called the letter ...very weak. *(WSJ2308)*

(without UID)	$P'_s(exp\|s, C) = 0.813$	$P'_s(imp\|s, C) = 0.187$	
(with UID)	$P_s(exp\|s, C) = 0.535$	$P_s(imp\|s, C) = 0.465$	Prediction= Explicit

(7) Full-time residential programs ... are particularly expensive – more per participant than a year at Stanford or Yale. (Implicit:**but**; COMPARISON.CONTRAST) Non-residential programs are cheaper, ... *(WSJ2412)*

(without UID)	$P'_s(exp\|s, C) = 0.649$	$P'_s(imp\|s, C) = 0.351$	
(with UID)	$P_s(exp\|s, C) = 0.383$	$P_s(imp\|s, C) = 0.617$	Prediction= Implicit

In Examples 5 and 6, the UID bias does not affect the prediction based on DC informativeness alone, since the CONTINGENCY.CAUSE.REASON sense is dominantly implicit (Example 5) and the COMPARISON.CONTRAST.OPPOSITION sense is dominantly explicit (Example 6), according to the probability distribution in the corpus. In these cases, argument informativeness has little effect on the RSA model. In contrast, in Example 7, the COMPARISON.CONTRAST sense could be expressed explicitly or implicitly, and the UID bias reverses the prediction based on DC informativeness. The model predicts that the speakers would not over-specify the discourse relation with a DC, since there are enough informative signals in the arguments (e.g., *expensive* vs. *cheaper* or *residential* vs. *non-residential*).

However, our approach, which approximates the argument informativeness based on the probability output of an automatic discourse parser, is limited by the accuracy of the discourse parser, as shown in Example 8.

(8) "Jeux Sans Frontieres"... is a hit in France. (Implicit:**but**; COMPARISON.CONTRAST)
A U.S.-made imitation under the title "Almost Anything Goes" flopped fast. *(WSJ2361)*

(without UID)	$P'_s(exp\|s, C) = 0.762$	$P'_s(imp\|s, C) = 0.238$	
(with UID)	$P_s(exp\|s, C) = 0.624$	$P_s(imp\|s, C) = 0.376$	Prediction= Explicit

The parser detects low informativeness in the arguments, and thus the model wrongly predicts that explicit marking is more likely. A possible explanation for this is that the constrast between *hit* and *flopped* is uncommon, and the parser fails to identify it as a discourse-informative signal. The performance of the discourse parser we used in the experiment is not yet satisfactory. The accuracy of the marking model could be improved with a more accurate discourse parser. In addition, the classifier of the discourse parser may have poorly calibrated probabilities, which means the probability estimates of the parser may not be well associated with how well the parser detects discourse signals. The association, and thus the overall performance of the model, may be improved by an additional probabilistic calibration step on the parser output (Nguyen and O'Connor, 2015).

Lastly, we discuss the potential to use behavioral experiments to further validate or improve the proposed method. The experiment described in Section 4 evaluates the prediction ability of the model against the actual data in the PDTB. In other words, the marking of each discourse relation chosen by the writers of the *Wall Street Journal* and the label attributed by PDTB annotators are taken as the gold standard. However, it is possible that other writers would choose differently, given the same relation sense and context. A behavioral experiment could be carried out to compare the judgment of multiple human speakers with the judgment of the annotators of PDTB and writers of the *Wall Street Journal*, as well as with the model's predictions. For example, Patterson and Kehler (2013) use a readability judgment task to test speakers' choice of explicit or implicit DCs in PDTB. They find that only 68% of the human judgment match the actual data in the PDTB, suggesting a high level of optionality in marking preference.

However, readability judgment may be biased towards explicit marking because it is generally agreed that explicit DCs facilitate discourse relation comprehension (Britton et al., 1982; Haberlandt, 1982; Kamalski et al., 2008; Loman and Mayer, 1983; Lorch Jr and Lorch, 1986; Meyer

et al., 1980; Millis and Just, 1994; Sanders et al., 1992; Sanders and Noordman, 2000). The marking of discourse relations could be examined using a production-oriented experimental design, such as the picture-to-language transcription task described in Soria and Ferrari (1998). Another option is to use a cloze test, in which subjects are presented with the discourse arguments and relation sense to be conveyed and are asked to fill in a DC or leave the relation implicit. Following the recent success in crowdsourced discourse annotation (Rohde et al., 2015; Scholman et al., 2016), a large number of judgments per relation can be collected by crowdsourcing such that a distribution of the marking preference can be obtained.

6. Conclusion

We have presented a language production model that predicts whether speakers will choose to use an explicit DC or not given the discourse relation they want to express. Our model gives a cognitive account of the speakers' choice and its results outperform those of previous work on the same task.

Although the option of DC marking is a subtle preference in the absence of other grammatical constraints, our proposed marking model tackles the option as a rational preference by the speaker. Using an information-theoretic approach, the model predicts a speaker's choice by balancing the advantage (informativeness) and disadvantages (production cost and redundancy) of using an explicit marker.

This is the first work to apply the RSA framework to discourse processing. Based on the discourse context, we adjusted the universal distribution of utterances and senses. This work also practically incorporates the UID principle, using the output of a discourse parser. We plan to expand the model to discriminate the choice of different explicit DCs and assess the effectiveness of the model in applications such as natural language generation or machine translation tasks, where models trained on monolingual data of different languages are applied. The long-term goal is to make use of the *listener model* to design a full, incremental discourse parsing algorithm that is motivated by the psycholinguistic reality of human discourse processing.

References

David Allbritton and Johanna Moore. Discourse cues in narrative text: Using production to predict comprehension. In *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

Fatemeh Torabi Asr and Vera Demberg. Implicitness of discourse relations. In *Proceedings of the International Conference on Computational Linguistics*, pages 2669–2684. Citeseer, 2012.

Fatemeh Torabi Asr and Vera Demberg. On the information conveyed by discourse markers. In *Proceedings of the Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 84–93, 2013.

Fatemeh Torabi Asr and Vera Demberg. Uniform information density at the level of discourse relations: Negation markers and discourse connective omission. *Proceedings of the International Conference on Computation Semantics*, pages 118–128, 2015.

Matthew Aylett and Alice Turk. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56, 2004.

Peter Baumann, Brady Clark, and Stefan Kaufmann. Overspecification and the cost of pragmatic reasoning about referring expressions. *Proceedings of the Annual Conference of the Cognitive Science Society*, 2014.

Anton Benz, Gerhard Jäger, and Robert Van Rooij. *Game theory and pragmatics*. Springer, 2016.

Leon Bergen, Roger Levy, and Noah D Goodman. Pragmatic reasoning through semantic inference. *Unpublished manuscript*, 2014.

Or Biran and Kathleen McKeown. Discourse planning with an n-gram model of relations. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 1973–1977, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

J Kathryn Bock. Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387, 1986.

Kathryn Bock, Gary S Dell, Franklin Chang, and Kristine H Onishi. Persistent structural priming from language comprehension to language production. *Cognition*, 104(3):437–458, 2007.

Bruce K Britton, Shawn M Glynn, Bonnie J Meyer, and MJ Penland. Effects of text structure on use of cognitive capacity during reading. *Journal of Educational Psychology*, 74(1):51, 1982.

Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. RST discourse treebank, 2002.

Herbert H Clark. Using language. *Journal of Linguistics*, 35(1):167–222, 1999.

Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

Laurence Danlos. Linguistic ways for expressing a discourse relation in a lexicalized text generation system. *Workshop of Discourse Relations and Discourse Markers*, pages 50–53, 1998.

Debopam Das, Maite Taboada, and Paul McFetridge. RST signalling corpus, 2015.

Paul E Engelhardt, Karl GD Bailey, and Fernanda Ferreira. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573, 2006.

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

Austin Frank and T Florian Jaeger. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 933–938, 2008.

Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998, 2012.

Albert Gatt, Roger PG van Gompel, Kees van Deemter, and Emiel Krahmer. Are we Bayesian referring expression generators. In *Proceedings of the Workshop on Production of Referring Expressions: Bridging the Gap between Cognitive and Computational Approaches to Reference*, 2013.

Dmitriy Genzel and Eugene Charniak. Entropy rage constancy in text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 199–206, 2002.

Noah D Goodman and Daniel Lassiter. *Probabilistic semantics and pragmatics: Uncertainty in language and thought*. Wiley-Blackwell, 2014.

Noah D Goodman and Andreas Stuhlmüller. Knowledge and implicature: modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184, 2013.

H Paul Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.

Stefan Th Gries. Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, 34(4):365–399, 2005.

Brigitte Grote and Manfred Stede. Discourse marker choice in sentence planning. In *Proceedings of the International Workshop on Natural Language Generation*, pages 128–137, 1998.

Karl Haberlandt. Reader expectations in text comprehension. *Advances in Psychology*, 9:239–249, 1982.

Jet Hoek and Sandrine Zufferey. Factors influencing the implicitation of discourse relations across languages. In *Proceedings the Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 39–45. TiCC, Tilburg center for Cognition and Communication, 2015.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted JM Sanders. The role of expectedness in the implicitation and explicitation of discourse relations. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 41–46. Association for Computational Linguistics, 2015.

T Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62, 2010.

T Florian Jaeger and Neal Snider. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, page 827, 2008.

Gerhard Jäger. *Game theory in semantics and pragmatics*, volume 3, pages 2487–2425. Mouton de Gruyter, 2012.

Lifeng Jin and Marie-Catherine de Marneffe. The overall markedness of discourse relations. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 1114–1119, 2015.

Judith Kamalski, Ted Sanders, and Leo Lentz. Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, 45(4-5): 323–345, 2008.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007, 2014.

James M Kilner, Karl J Friston, and Chris D Frith. Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 8(3):159–166, 2007.

Gina R Kuperberg, Martin Paczynski, and Tali Ditman. Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, 23(5):1230–1246, 2011.

Daniel Lassiter and Noah D Goodman. Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and Linguistic Theory*, 23:587–610, 2013.

Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, pages 1–36, 2015.

Willem JM Levelt and Stephanie Kelter. Surface form and memory in question answering. *Cognitive psychology*, 14(1):78–106, 1982.

Stephen C Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press, 2000.

Roger Levy and T. Florian Jaeger. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, (849-856), 2006.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 283–288, 2014.

Ziheng Lin, Minyen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 343–351, 2009.

Nancy L Loman and Richard E Mayer. Signaling techniques that increase the understandability of expository prose. *Journal of Educational psychology*, 75(3):402, 1983.

Robert F Lorch Jr and Elizabeth Pugzles Lorch. On-line processing of summary and importance signals in reading. *Discourse Processes*, 9(4):489–496, 1986.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkey, Steven J. Bethard, and David McClosky. The Standord CoreNLP natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

Bonnie JF Meyer, David M Brandt, and George J Bluth. Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading research quarterly*, pages 72–103, 1980.

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

Keith K Millis and Marcel Adam Just. The influence of connectives on sentence comprehension. *Journal of Memory and Language*, 33(1):128–147, 1994.

Will Monroe and Christopher Potts. Learning in the Rational Speech Acts model. *arXiv preprint arXiv:1510.06807*, 2015.

Megan Moser and Johanna D Moore. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 130–135. Association for Computational Linguistics, 1995.

John D Murray. Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25 (2):227–236, 1997.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598, 2015.

Naho Orita, Eliana Vornov, Naomi H. Feldman, and Hal Daumé III. Why discourse affects speakers' choice of refering expressions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1639–1649, 2015.

Joonsuk Park and Claire Cardi. Improving implicit discourse relation recognition through feature set optimization. *Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112, 2012.

Gary Patterson and Andrew Kehler. Predicting the presence of discourse connectives. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 914–923, 2013.

Martin J Pickering and Simon Garrod. Do people use language production to make predictions during comprehension? *Trends in cognitive sciences*, 11(3):105–110, 2007.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. Easily identifiable discourse relations. Technical report, University of Pennsylvania, 2008.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 683–691, 2009.

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. Manuscript, 2015.

Rashmi Prasad, Nikhit Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. *Proceedings of the Language Resource and Evaluation Conference*, pages 2961–2968, 2008.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, pages 921–950, 2014.

Hannah Rohde and William S Horton. Anticipatory looks reveal expectations about discourse relations. *Cognition*, 133(3):667–691, 2014.

Hannah Rohde, Anna Dickinson, Chris Clark, Annie Louis, and Bonnie Webber. Recovering discourse relations: Varying influence of discourse adverbials. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, page 22, 2015.

Attapol Rutherford and Nianwen Xue. Discovering implicit discourse relations through Brown cluster pair representation and coreference patterns. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, 2014.

Ted Sanders. Coherence, causality and cognitive complexity in discourse. In *Proceedings of the Symposium on the Exploration and Modelling of Meaning*, 2005.

Ted JM Sanders and Leo GM Noordman. The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29(1):37–60, 2000.

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35, 1992.

Merel CJ Scholman, Jacqueline Evers-Vermeul, and Ted JM Sanders. Categories of coherence relations in discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, 7(2):1–28, 2016.

Donia Scott and Clarisse Sieckenius de Souza. Getting the message across in RST-based text generation. *Current research in natural language generation*, 4:47–73, 1990.

Erwin M Segal, Judith F Duchan, and Paula J Scott. The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse processes*, 14(1): 27–54, 1991.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (379-423; 623-656), 1948.

Mark Smith and Linda Wheeldon. Syntactic priming in spoken sentence production–an online study. *Cognition*, 78(2):123–164, 2001.

Claudia Soria and Giacomo Ferrari. Lexical marking of discourse relations-some experimental findings. In *Proceedings of the ACL-98 Workshop on Discourse Relations and Discourse Markers*, pages 36–42, 1998.

Harry Tily and Steven Piantadosi. Refer efficiently: Use less informative expressions for more predictable meanings. *Proceedings of the workshop on the production of referring expressions*, 2009.

Jianxiang Wang and Man Lan. A refined end-to-end discourse parser. *CoNLL 2015*, pages 17–24, 2015.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational linguistics*, 29(4):545–587, 2003.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. The Discourse Graphbank: A database of texts annotated with coherence relations, 2005.

Frances Yung, Kevin Duh, and Yuji Matsumoto. Crosslingual annotation and analysis of implicit discourse connectives for machine translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 142–152, 2015.

Frances Yung, Kevin Duh, Taku Komura, and Yuji Matsumoto. Modeling the interpretation of discourse connectives by Bayesian pragmatics. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 531–536, 2016.

Henk Zeevat. Bayesian interpretation and optimality theory. *Bidirectional Optimality Theory. Palgrave Macmillan, Amsterdam*, pages 191–220, 2011.

Henk Zeevat. *Perspectives on Bayesian Natural Language Semantics and Pragmatics*, pages 1–24. Springer, 2015.