CLIPS: A Chinese Legal Corpus with Discourse Annotations

Hong Wang[1,2] and Yunfeng Ge[1]✉

[1] Foreign Languages College, Shandong Normal University

Jinan 250014, Shandong, China

[2] National Key Research Center for Linguistics & Applied Linguistics,

Guangdong University of Foreign Studies, Guangzhou 510420, Guangdong, China

wanghonghelen@126.com, dennisge@126.com

Abstract

The paper presents CLIPS (the CLIPS), a corpus annotated with discourse relations based on Discourse Information Theory (DIT) that takes account of both macro- and micro- structures at discourse level. The information units and information elements in Chinese legal discourses are firstly characterized, and then a 16-valued classification of information units for macro-structure of discourse relations and a 25-valued classification of information elements for micro-structure of discourse relations are introduced. The paper also describes how the annotation strategy procedure is designed and the annotation conduction based on the above characterization.

Keywords

discourse annotation, information unit, information knot, information element, info-tagging

1. Introduction

In the realm of Natural Language Processing (NLP), researches have been conducted on various linguistic levels, extending from lexical, syntactic, semantic, and pragmatic to discursive. NLP studies at discourse level are aimed at handling with the overall structure of discourses that ex-

tend across sentential boundaries and serving for researches on both corpus linguistics and discourse analyses. The present field of NLP and computational linguistics is dominated by statistical and data-driven approaches; hence a consistently annotated linguistic corpus is an indispensable resource that promotes the advances in particular areas (Zhou & Xue, 2015). Annotation becomes a necessary prerequisite to subsequent applications of linguistic corpus, such as linguistic information extraction, and so on.

Most of the existing annotations at discourse level have been focused on English, among which the Penn Discourse Treebank (PDTB-Group, 2008) and the RST Discourse Treebank (Carlson, Marcu, & Okurowski, 2003) represent corpora that are conducted on the basis of various linguistic components, including discourse connectives, grammatical rules, and independent clauses. Most discourse annotation studies on Chinese followed either PDTB or the RST Discourse Treebank, which still have problems in expanding the range of linguistic components and in handling with some large-size units of texts (Marcu, Carlson, & Watanabe, 2000), like independent sentences and recognizing relations between sentences. This article will describe the annotation process of a Chinese legal corpus that is annotated with discourse information, and it is hoped to spur the advances of Chinese annotation at discourse level, among its other possible uses.

Being based on a well-grounded particular linguistic theory helps the creation of a corpus in its operationalization and in its scaling up to domain-independent discourses (Zhou & Xue, 2015). The present research adopted Discourse Information Theory (DIT) (Du, 2014, 2007; Ge, 2018, 2016; Wang & Ge, 2016) as theoretical basis, which defines discourse information units as basic components of a discourse and discourse information elements as essential components of a discourse information unit. DIT is used as the theoretical basis in the present study for three reasons:

- It is a framework that can uniformly capture cognitive, functional, and discursive features of a discourse.
- It gives consideration to both the overall structure and fine details of a discourse; hence it can present a discourse stereoscopically.

- It provides a comprehensive and effective classification of discourse information for Chinese discourses.

In the DIT theoretical framework (Du, 2014), the integral structure of a discourse can be represented as a tree defined in terms of six aspects:

- Each discourse contains a topic, which is defined as kernel proposition (KN). Information units are the basic components of a discourse. Information elements compose an information unit.
- As the principal line of a discourse, information trunk is based on KN, around which discourse information develops and flows.
- The major structural patterns of information trunk consist of 4 types, including causal pattern, enumerating pattern, parallel pattern, and subordinative pattern.
- Information branches function as a transition from trunk to leaves. Branches are constrained by trunk and constrain developments of leaves as well.
- The 15 types of information knots (i.e. the internal nodes of the tree) correspond to various relations that hold between two adjacent information units.
- Information leaves are the terminal of information branches. Density of information leaves signifies development degree of discourse information.

2. Building up discourse structure

The first step in characterizing the structure of a discourse is to determine the elementary units of discourse, which are the building blocks of information tree. According to DIT (Ge, 2018; Du, 2014, 2007), information is defined as the minimal integral meaningful unit in communication. The present study follows DIT to choose information unit as the elementary constituent of discourse and information element as basic component of information unit.

Since information units are defined as the elementary units of a discourse, the discourse relation between two adjacent information units is indicated via information knots, all information knots are connected together to create a hierarchical structure, which is the visualized inter- and intra-sentential relations in a discourse.

The analysis and annotation of a discourse are conducted at the following two different levels.

- *Hierarchical*: Information units in a discourse are divided into different levels according to their distance to *KN*. Level-1 information units explain the *KN* directly, and each level-1 information unit consists of many level-2 information units, which explains their direct upper-level information unit. Similarly, a level-2 information unit is surrounded by its level-3 information units as elaborations. The relations between the above three levels are hierarchical.
- *Horizontal*: Relations information units at the same level are horizontal. Likewise, relations between information elements in a same information unit are horizontal.

In addition to *KN*, the topic of a discourse, there are 15 types of information knots that can be used in annotating different relations between information units.

- *WT*: General formulation of things, including objects, events, and behavior. Example (1) is a statement of the close of a case.

(1) 本案 现已 审理 终结。
Ben an xian yi shenli zhongjie.
The case has now come to a close.

- *WB*: Basis for handling issues, including laws and reference standards. Often indicated by "根据(according to)", *WB* usually appears as a phrase but functions as a core part of the meaning of a sentence. In example (2), the phrase initiated by "根据" shows the basis of Company B's leasing equipment. Since the phrase occupies the core of the meaning, the information unit is classified as a WB.

(2) 根据 A 公司 2005 年 发布 的 公告， 自 2004 年 起, B 公司
Genjv A gongsi 2005 nian fabu de gonggao, zi 2004 nian qi, B gongsi
According to the announcement released by company A in 2005, from 2004, company B

租赁 A 公司 的 3 万吨 电解铝 设备。
zulin A gongsi de 3 wandun dianjielv shebei.

had leased 30,000-ton electrolytic aluminum equipment from company A.

- *WF*: Statement of underlying objective facts without any subjective inference. Indicative mood is usually employed. List of WF often comes with source or basis of information. Example (3) is an objective description of the fact that two companies signed 5 agreements.

(3) A 公司 和 B 公司 签署了 5 份 《债务转移协议》。
A gongsi he B gongsi qianshu le 5 fen zhaiwu zhuanyi xieyi.
Company A and company B has signed five Debt Transfer Agreements.

- *WI*: Inference about people or thing to show reasoning process. Descriptions of judgment are widely used. Most descriptions contain "是", "认为", "觉得", "应当", "也就是说". Example (4) explains reason through inferential process why Company A should bear the repayment liability.

(4) A 公司 应 依据《债务 转移 协议 补充 承诺》 承担 还款 责任。
A gongsi ying yijv zhaiwu zhuanyi xieyi buchong chengnuo chengdan huankuan zeren.
Company A should bear the repayment liability under the Supplementary Letter of Commitment to Debt Transfer Agreements.

- *WP*: Putting forward solution or suggested disposal to a problem. Not having distinct and unique expression like WI. Example (5) is the plaintiff's request of Company D's repayment.

(5) 原告 请求 判令 D 公司 对 上述 债务 在 其 担保 的 范围 内
Yuangao qingqiu panling D gongsi dui shangshu zhaiwu zai qi danbao de fanwei nei
The plaintiff requests the court to rule that Company D should assume the joint and several liability for

承担 连带 清偿 责任。
chengdan liandai qingchang zeren.
repaying the above debts within the extent of guarantee that it provided.

- *WO*: Usually being related to people, deictic expressions, and proper nouns. Common expressions include description of person(s). Example (6) is about persons who are involved in the quarrel.

(6) 被告人 刘吉森、 徐振博 酒后 因 琐事 与 被害人 张某 发生口角。
Beigaoren Liu Jisen、 Xu Zhenbo jiuhou yin suoshi yu beihairen Zhangmou fasheng koujiao.
The defendant Liu Jisen and Xu Zhenbo were drunk and quarreled with victim Zhang for trifles.

- *WN*: Including all concepts of time indicated by corresponding demonstrative words of time. Example (7) is about the time when Company B repaid two loans.

(7) B 公司 于 2016 年 8 月 5 日 分别 偿还 了 该 两笔 借款。
B gongsi yu 2016 nian 8 yue 5 ri fenbie changhuan le gai liangbi jiekuan.
Company B repaid these two loans respectively on Aug. 5, 2016.

- *WR*: Referring to place, including location, tendency, source, etc. Example (8) provides the appellant's detailed address.

(8) 上诉人 的 经常 居住地 为 天津市 南开区 密云路 川北里 3-1-501 号。
Shangsuren de jingchang jvzhudi wei tianjinshi nankaiqu miyunlu chuanbeili 3-1-501 hao.
The appellant's habitual residence is 3-1-501, Chuanbei Lane, Miyun Road, Nankai District, Tianjin City.

- *HW*: The way things go or problems being solved. Example (9) explains the way in which Company B purchased the use rights of some land of Company A. Similarly, example (10) is about how the loan principal comes into being.

(9) B 公司 以 承担 A 公司 所 欠 金融机构 收账款 及 现金支付 方式，
B gongsi yi chengda A gongsi suo qian jinrongjigou shouzhangkuan ji xianjinzhifu fangshi,
Company B purchased the use rights of some land of Company A

收购 A 公司 的 部分 土地 使用权。
shougou A gongsi de bufen tudi shiyongquan.

through assuming debts owed by Company A to financial institutions and cash payments.

(10) 本案 贷款 本金 是 2000 年 旧贷款 经 多次 以贷还贷 方式 演化 而来。
Ben'an daikuan benjin shi 2000 nian jiudaikuan jing duoci yidaihuandai fangshi yanhua erlai.
The loan principal in this case was loans gradually evolving from the old loans in 2000 through several times of repayment of old loans with new ones.

- *WY*: Showing reasons. The first two lines of example (11) are providing reasons to explain the reduction of Company B's property.
- *WE*: Explaining favorable and unfavorable effects of things. The third line of example (11) is the unfavorable effect of Company B's improper investment.

(11) B 公司 以其 优良 资产 与 他人 组建 A 公司，
B gongsi yi qi youliang zichan yu *taren zujian A gongsi,*
Company B established Company A jointly with others with its good assets,

将 净值 9000 万元 的 资产 投入到 A 公司，
jiang jingzhi 9000 wanyuan de zichan tourudao A gogsi,
and invested assets with a net value of 90 million yuan in Company A,

导致 其 偿还 银行 债务 的 责任 财产 减少。
daozhi qi changhuan yinhang zhaiwu de zeren caichan jianshao.
causing the reduction of its property with which it assumed the liability for repayment of bank debts.

- *WC*: Providing condition under which things exist or change. Example (12) is aimed at explaining the condition under which another page can be added.

(12) 经 书记员 同意 可以 另页 补正。
Jing shujiyuan tongyi keyi lingye buzheng.
Being agreed by the court clerk, another page can be added for supplement and correct.

HW, *WY*, *WE*, and *WC* can be expressed in various linguistic forms, including words, phrases, sentence fragments, clauses, and independent sentences.

- *WA*: Expressing people's attitude, appraisal, tendency, slanting opinions by employing general descriptions. In example (13), Sanmenxia Station Branch of ICBC declares its attitude by using the approval expression "予以认可".

(13) 三门峡车站工行对此予以认可。
Sanmenxia chezhann gonghang duici yuyi renke.
Sanmenxia Station Branch of ICBC recognized the above facts.

- *WG*: Showing condition contrast before and after a certain process. The first part in example (14) is a *WF* that shows the original stand of the court, while the second part is a *WG* for its showing the change of the court's stand.

(14) 原审 法院 已 支持 A 公司 的 主张， 却 又 认定 A 公司 应 承担 还款责任。
Yuanshen fayuan yi zhichi A gongsi de zhuzhang, que you rending A gongsi ying chengdan huankuan zeren.
The court of the original instance has already sustained company A's claim, but held that company A should assume the repayment liability.

- *WJ*: Providing conclusive argument or overall conclusion after reasoning and argumentation. Having highly-recapitulative force. Usually following a series of *WF*s and *WI*s. Both examples (15) and (16) are judgments of a court, in which (15) is the ruling and (16) is the overruling of the plaintiff's claims.

(15) 原审 法院 判决，B 公司 向 工行 偿还 3375 万元。
Yuanshen fayuan panjue, B gongsi xiang gonghang changhuan 3375 wanyuan.
The court of the original instance ruled that Company B should repay 3375 million yuan to ICBC.

(16) 驳回 三门峡 车站 工行 的 其他 诉讼 请求。
Bohui Sanmengxia chezhan gonghang de qita susong qingqiu.
Other claims of Station Branch of ICBC should be overruled.

Information knots are used to annotate inter-relations between information units, while information elements are used in annotating inner-relations of the components in a same infor-

mation unit. Information elements are partitioned into 3 classes with 25 kinds of different values as shown in Table 1. The 3 classes include Process, Entity, and Condition, each of which has different amount of values respectively to indicate various aspects of each class.

Table 1. Information elements tagset

Info Elements	**Value**	**Symbol**	**Meaning**
Process	State	S	The existence and representation of things.
	Quality	Q	Judgment of characteristics of things.
	Appear	A	Conjuring things from nothing.
	Relation	R	Connection between things.
	Behavior	B	Activities of creatures or things.
	Cause	C	Things that bring others into existence.
	Turn	T	Changes in existence, nature, or activity.
	Negation	N	Negative elements.
Entity	Agent	n	The doer of an action
	Dative	d	The creature influenced by an action or state.
	Patient	p	The passive person.
	Factitive	f	Person or thing being created.
	Attribute	b	Explanation of another individual.
Condition	Instrument	i	Substance being used in an action.
	Location	l	Physical or abstract position.
	Source	s	The place where things come from.
	Goal	g	Destination, goal.
	Comititive	c	Things come into being or change along with the process.
	Time	t	Time point or time quantum.
	Affected	a	Influences exerted by things or actions.
	With	w	Measure or condition with which a process starts, proceeds or ends.
	Situation	o	Broader context.
	Basis	b	Things providing foundation.
	Manner	m	The way actions or things happen or proceed.
	Elaboration	e	Illustrations or details.

3. Discourse annotation task

3.1 Annotator profile and training

Since present research is aimed at building a high-quality corpus that is consistently annotated at discourse level, annotators were carefully selected and finally 2 PhD. candidates with academic experience in forensic linguistics and discourse analysis and 4 postgraduates majoring in applied linguistics were sorted out.

The PhD. candidates had been quite familiar with DIT and Info Tagging. The 4 postgraduates were firstly trained to get familiar with the theoretical basis DIT and the accessorial tool Info Tagging in about three months. During this period, the training was focused on recognizing and differentiating information units and information elements, and learning the mechanics of Info Tagging. In the fourth month, annotators were asked to annotate a same short discourse independently under the guidance of pre-set specifications. Then the 6 annotators met as a group to discuss the differences between the annotations of the 4 postgraduates', which range from ascertaining the scope of information units, classifying information units, and recognizing information knots, to categorizing information elements. Finally, they discussed ways to enhance the consistency of annotation.

3.2 Annotator tool—Info Tagging

Info Tagging is an accessorial annotation tool used in the discourse annotation process. Firstly, annotators need to determine the range of each information unit in a discourse and to divide information elements in each information unit. Subsequently, the Doc head and the main body of the discourse are annotated with the help of Info Tagging. By annotating a discourse, we can show the topic of the discourse, subtopics of each sentence in the discourse, hierarchical structure of the discourse, functional relations between sentences, and inner structure of each sentence.

Fig. 1 shows the process of Doc head annotation, which is a preprocessing of annotation that is aimed at case classification and case retrieval in CLIPS. Doc head annotation includes spotting the *KN* and completing kinds of peripheral information.

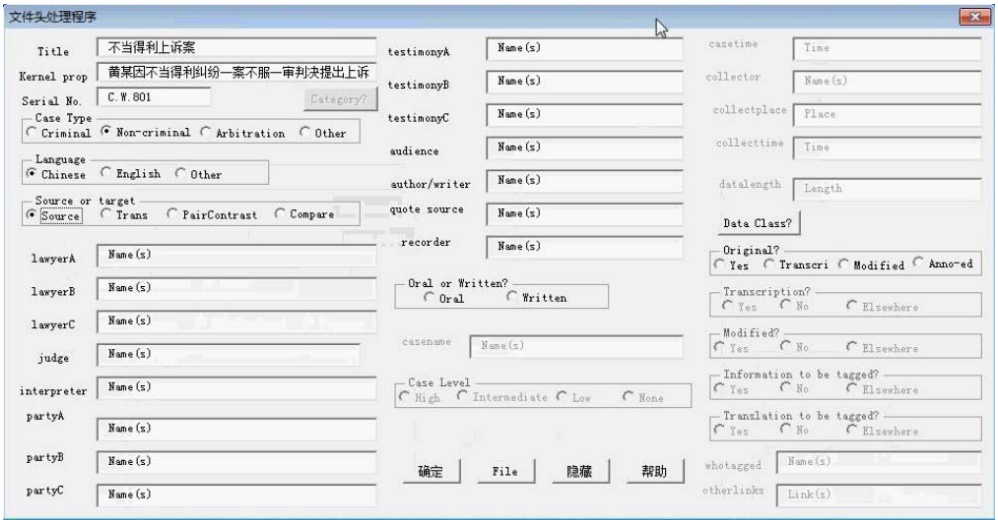

Figure 1. Doc head annotation.

The *Title* box displays the topic (i.e. contents of Doc head) of a discourse, that is, "Appeal on unjust enrichment (不当得利上诉案)", which makes for later information retrieval. The *Kernel prop* box functions as the topic of the whole discourse, which includes the concerned person(s), event(s), and reason(s). The *Serial No.* box shows the sequential number of a case in CLIPS. *Case Type* includes four choices, namely, criminal, non-criminal, arbitration, and others. *Language* involves the language used in the data, which mainly are Chinese and English at the moment. *Source or target* includes four choices, which respectively stand for language situation of the original discourse. The following boxes include concrete names of *lawyer (A, B, C)*, *judge*, *interpreter*, *party (A, B, C)*, *testimony (A, B, C)*, *audience*, *author*/*writer*, *quote source*, and *recorder*. *Oral or Written* shows whether the discourse is transcribed. *Casename* and *Case Level* are about the specific name of a case and level of the court in which the case was heard. There are 5 other boxes for case details, including *casetime*, *collector*, *collectplace*, *collecttime*, and *datalength*, which exhibit time of the case, name(s) of information collector, place and time of discourse collection, and length of the discourse. Details of data in the lower right corner

show whether the data have been operated, to be specific, whether the data is original or transcribed or modified or annotated.

After having processed the topic and details of the discourse, annotators begin to deal with the main body of the discourse. The main body is divided into several parts and subtopic of each part is summarized; and then, the *hierarchical* level of each information unit is analyzed.

Figure 2. Main body annotation.

The panel in Fig. 2 shows the process of their analysis, which gives consideration to both subtopic and hierarchical relations between an information unit and its upper level unit. On this panel, *ParentInfo* (父信息) shows the upper level that the information unit subordinate to. *ChildInfo* (子信息) exhibits the information unit at its own level. *InfoKnot* (信息点) represents the discourse relation both hierarchically and horizontally between two information units. *InfoValue* (信息点值) tells from whom the information unit comes, including lawyers, either of the two parties, and testimonies. *KeyWords* (关键词) shows the extracted topic of the information unit. *Result* (选择结果) completely presents the annotation contents of an information unit. With this annotation result, it is easily to recognize the subtopic of each information unit and hierarchical relations between information units. Hence, annotations can be presented as a tree diagram, exhibiting topic and relations contained in the whole discourse and how information flows in the discourse.

Fig. 3 shows a fragment of an annotated Chinese civil trial transcript. There are four information knots in Fig. 3, which are two *WT*s and two *WP*s. In the annotation, the serial numbers *<1,2,2,8>* shows the hierarchical order of the information knot, that is, this *WP* is the 8th level-2 information knot that is subject to the 2nd level-1 information knot. The letters *P,O* refers to the value of information knot is positive. Letter *J* indicates the information source. Letter *Z* shows there's no quotation in this information knot. Letter *A* is an information sharing category, which implies that the information knot is known to A, but not B. The numbers *<0,0,0,0>* shows there's no repetition in this information knot.

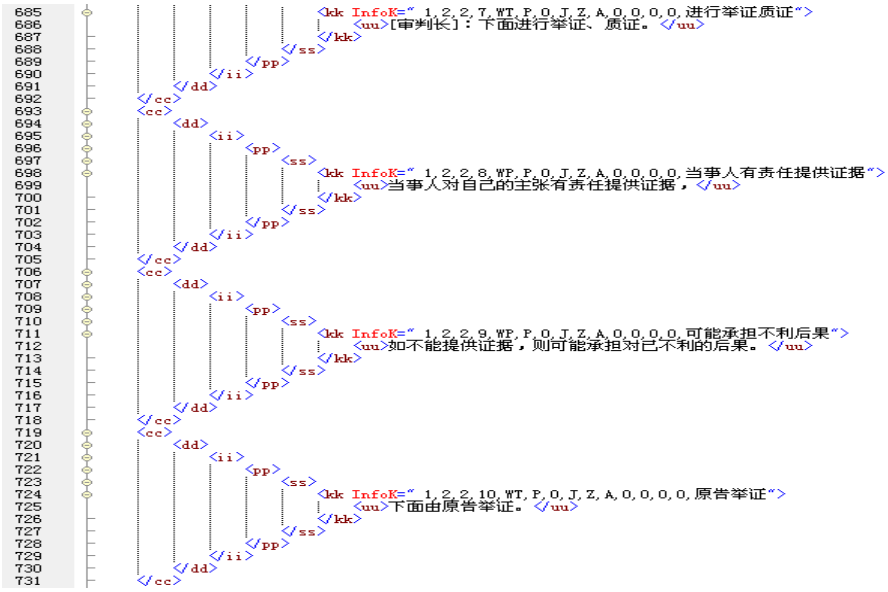

Figure 3. Fragment of an annotated discourse

4. Corpus statistics, evaluation, and application

4.1. Corpus statistics

The CLIPS consists of legislative documents, judicial judgments, transcripts of judicial trials, and law dissemination discourses, totally representing over 160 million Chinese characters.

In selecting these documents, we referred to Chinese laws and regulations, trials from courts in Guangzhou, Beijing, and He'nan.

CLIPS consists of different sub-corpora, i.e. the core part, the ordinary part, the communicative part, and the outside part. Among which the core part of CLIPS has been annotated with discourse information grounded on elaborated analysis.

The annotated corpus can display hierarchical information structure of a whole discourse, including contents of information units, symbols of information knots, location of information knots in both hierarchical and horizontal structures, parent and child information, and so on.

4.2 Annotation evaluation

During the annotation process, each discourse in the core part of CLIPS is double-annotated by those postgraduate annotators independently, and then checked by another more experienced annotator. The inter-annotator agreement is computed for evaluation by using accuracy, precision, recall, F1-score, and Kappa.

Based on exact match, the calculation of accuracy is computed as the number of matched recognition of information units between two annotators divided by the total number of information units both of the two annotators recognized. In this way, the accuracy is 98.3%, and the performance of the annotation reflects good agreement (with P=94.8%, R=97.6%, F1-score=96.5%, and kappa=0.72).

4.3 Application of CLIPS

The annotated data in CLIPS have already served forensic authorship attribution, forensic speaker recognition, Chinese-English translation, and psychological experiments.

Many features of discourse information extracted from the annotated data in CLIPS have been effectively used in authorship attribution of Chinese emails (Zhang, 2016), recognizing speakers (Guan, 2016), and evaluating the reliability of witness testimony for lie detection (Yu, 2017), including number of objective and subjective information units, ratio of subjective to objective units, number of information elements per information unit, number of different information unit and information functions, which can efficiently profile the unique features of a

certain type of discourse, detect similarities and differences between a certain number of discourses.

5. Conclusion

The present research introduced the Chinese corpus of legal discourses CLIPS, which is annotated at discourse level by adopting DIT that take the form of information units and information elements. Different from other existing discourse corpora, annotation of CLIPS took into account both the relations between sentences in a discourse and fine details in each sentence. The annotation took into consideration the deficiency of ordinary sentence segmentation and hence segmented each discourse according to both integral meaning expression and punctuation.

The main feature of the annotation system is that it can clearly mark both the overall information structure of a discourse and the inner details of each information unit in the discourse, including the topic of the whole discourse, the number of information units at different levels in a discourse, the content and property of an information unit, the relations between information units, the hierarchical and horizontal level that an information unit belongs to, the constituents that an information unit contains, the property of these constituents, and the relations between constituents in a same information unit.

The annotation output of CLIPS can be presented both in XML format, in tree diagram, and in word document. Having put the DIT theory into application, CLIPS is anticipated to be multifunctional in linguistic researches and can support a wide range of applications in forensic linguistic studies, natural language processing and language technologies.

6. Acknowledgement

This study is supported by Project of the National Social Science Foundation via Grant No. 16BYY064, Project of the National Language Committee via Grant No. YB135-5, Project of Shandong Social Science Plan via Grant No. 13CWXJ23, and SDNU 2014 Young Teachers' Research Project.

7. References

Carlson, L., Marcu, D., & Okurowski, M. E. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current and New Directions in Discourse and Dialogue*, Kuppevelt, J. van and Smith, R. W. (Eds.). Heidelberg: Springer, pp. 85-112.

Marcu, D., Carlson, L., & Watanabe. M. 2000. The Automatic Translation of Discourse Structures. In *Proc. of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, pp. 9–17. Doi: 10.1.1.42.2061.

Du, J. 2014. *Studies on Legal Discourse Information.* Beijing: People's Publishing House.

Du, J. 2007. Study on Legal Discourse Tree Information Structure. *Modern Foreign Languages*. 30(1): 40-50.

Ge, Y. 2018. *Resolution of Conflict of Interest in Chinese Civil Court Hearings: A Perspective of Discourse Information Theory*. Bern: Peter Lang.

Ge, Y. 2016. Sensationalism in media discourse: A genre-based analysis of Chinese legal news reports. *Discourse & Communication.* 10(1): 22-39. Doi: 10.1177/ 1750481315602395.

Guan, X. 2016. Study on the effectiveness of ideolect in speaker recognization—From a perspective of discourse information. Journal of Zhongyuan University of Technology. 27(5): 14-18.

PDTB-Group. 2008. The Penn discourse Treebank 2.0 Annotation Manual. *Technical Report IRCS-08-01*.

Wang, H. & Ge, Y. 2016. Annotation scheme for legal discourse information and hierarchical levels. *Proc. of The 20th International Conference on Asian Language Processing (IALP)*. pp. 53-58. Doi: 10.1109/IALP.2016.7875933.

Yu, X. 2017. Evaluation of the Reliability of Witness Testimony on the Basis of Discourse Information Analysis. Unpublished doctoral thesis of Guangdong University of Foreign Studies.

Zhang, S. 2016. Authorship attribution and feature testing for short Chinese emails. The International Journal of Speech, Language and the Law. 23: 71-97.

Zhou, Y. & Xue, N. 2015. The Chinese Discourse TreeBank: a Chinese Corpus Annotated with Discourse Relations. *Language Resources and Evaluation*, 49(5): 397-431. Doi: 10.1007/s10579-014-9290-3.

* Yunfeng Ge is the corresponding author.