Discursive Usage of Six Chinese Punctuation Marks

YUE Ming
Department of Applied Linguistics
Communication University of China
100024 Beijing, China
yueming@cuc.edu.cn

Abstract

Both rhetorical structure and punctuation have been helpful in discourse processing. Based on a corpus annotation project, this paper reports the discursive usage of 6 Chinese punctuation marks in news commentary texts: Colon, Dash, Ellipsis, Exclamation Mark, Question Mark, and Semicolon. The rhetorical patterns of these marks are compared against patterns around cue phrases in general. Results show that these Chinese punctuation marks, though fewer in number than cue phrases, are easy to identify, have strong correlation with certain relations, and can be used as distinctive indicators of nuclearity in Chinese texts.

1 Introduction

Rhetorical structure has been proven useful in NLP projects such as text generation, summarization, machine translation and essay scoring. Automatic discourse parsing remains an elusive task, however, despite much rule-based research on lexical cues such as anaphora and conjunctions. Parsing through machine learning has encountered a bottleneck, due to limited resources--there is only one English RST treebank publicly available, and one RST-annotated German corpus on its way.

Punctuation marks (PMs) have been proven useful in RST annotation as well as in many other NLP tasks such as Part-of-Speech tagging, Word Sense Disambiguation, Near-duplicate detection, bilingual alignment (e.g. Chuang and Yeh, 2005), etc. Dale (1991) noticed the role of PMs in determining rhetorical relations. Say (1998) did a study on their roles in English discourse structure. Marcu (1997) and Corston-Oliver (1998) based their automatic discourse parser partially on PMs and other orthographical cues. Tsou et al. (1999) and Chan et al. (2000) use PMs to disambiguate candidate Discourse Markers for a Chinese summarization system. Reitter (2003) also used PMs to distinguish ATTRIBUTION and ELABORATION relations in his Feature-rich SVM rhetorical analysis system.

All these inspired us to survey on the rhetorical patterns around Chinese PMs, so as to provide more direct a priori scores for the coarse rhetorical analyzer by Zhang et al. (2000) in their hybrid summarization system.

This paper is organized into 5 parts: Section 2 gives an overview of a Chinese RST treebank under construction, and a survey on the syntax of six main PMs in the corpus: Colon, Dash, Ellipses, Exclamation Mark, Question Mark, and Semicolon. Section 3 reports rhetorical patterns around these PMs. Section 4 is a discussion on the effectiveness of these PMs in comparison with Chinese cue phrases. Section 5 is a summary and Section 6 directions for future work.

2 Overview of Chinese RST treebank under construction

2.1 Corpus data

For the purpose of language engineering and linguistic investigation, we are constructing a Chinese corpus comparable to the English WSJ-RST treebank and the German Potsdam Commentary Corpus (Carlson et al. 2003; Stede 2004). Texts in our corpus were downloaded from the official website of *People's Daily*[1], where important *Caijingpinlun*[2] (CJPL) articles

[1] www.people.com.cn.

[2] *Caijinpinglun* (CJPL) in Chinese means "financial and business commentary", and usually covers various topics in social economic life, such as fiscal policies, financial reports,

Proceedings of the COLING/ACL 2006 Student Research Workshop, pages 43–48,
Sydney, July 2006.

by major media entities were republished. With over 400 authors and editors involved, our texts can be regarded as a good indicator of the general use of Chinese by Mainland native speakers.

At the moment our CJPL corpus has a total of 395 texts, 785,045 characters, and 84,182 punctuation marks (including pruned spaces). Although on average there are 9.3 characters between every two marks, sentences in CJPL are long, with 51.8 characters per common sentence delimiters (Full Stop, Question Mark and Exclamation Mark).

2.2 Segmentation

We are informed of the German Potsdam Commentary Corpus construction, in which they (Reitter 2003) designed a program for automatic segmentation at clausal level after each Sign="$."(including {., ?, !, ;, :, …}) and Sign="$,"(including {,})[3]. Human interference with the segmentation results was not allowed, but annotators could retie over-segmented bits by using the JOINT relation.

Given the workload of discourse annotation, we decided to design a similar segmentation program. So we first normalized different encoding systems and variants of PMs (e.g. Dashes and Ellipses of various lengths), and then conducted a survey on the distribution (Fig. 1) and syntax of major Chinese punctuation marks (e.g. syntax of Chinese Dash in Table 1).

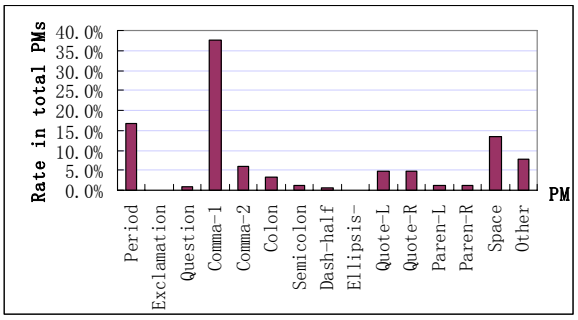

Figure 1: Percentage of major punctuation marks in the Chinese corpus[4]

C-Comma-1 is the most frequently used PM in the Chinese corpus. While it does delimit clauses, a study on 200 randomly selected C-Comma-1 tokens in our corpus shows that 55 of them are used after an independent NP or discourse marker. This rate, times the total number of C-Comma-1, means we would have to retie a huge number of over-segmented elements. So we decided not to take C-Comma-1 as a delimiter of our Elementary Unit of Discourse Analysis (EUDA) for the present.

Structure of C-———[5]	%
[NP+——NP+]NP	3.12%
[s+——s+]NP	0.44%
S*[NP——NP——VP]S	1.78%
S*[NP——s——VP]S	0.89%
S*[s——s——s]S	6.22%
<title>s+——Source：s+</title>	2.67%
<title>Source：s——s+</title>	0.44%
<para>S*s——</para>	1.33%
<para>S——S+</para>	2.22%
<para>S*s”——s+</para>	7.56%
<para>——S+</para>	12.44%
<para>S*s——s+</para>	60.89%
TTL	**100.00%**

Table 1: Syntax of Chinese Dash

42.9% of the colons in CJPL are used in the structural elements[6] of the texts. Other than these, 56.5% of the colons are used between clausal strings, only 0.6% of the colons are used after non-clausal strings.

99.6% instances of Exclamation Mark, Question Mark, Dash, Ellipses and Semicolon in the Chinese corpus are used after clausal strings.

In our corpus, 4.3% of the left quotation marks do not have a right match to indicate the end of a quote. Because many articles do not give clear indications of direct or indirect quotes[7], it is very difficult for the annotator to makeup.

Parentheses and brackets have a similar problem, with 3.2% marks missing their matches.

trading, management, economic conferences, transportation, entertainment, education, etc.

Collected by professional editors, most texts in our corpus are commentaries; some are of marginal genres by the Chinese standards.

[3] Dash, as a Sign= "$(", was not selected as a unit delimiter in the Potsdam Commentary Corpus.

[4] PMs are counted by individual symbols.

[5] The symbol "S" donates sentences with a common end mark, while "s" denotes structures orthographically end with one of the PMs studied here. "+" means one or more occurrences, "*" means zero or more occurrences. The category after a bracket pair indicates the syntactic role played by the unit enclosed, for example "[……]NP" means the ellipses functions as an NP within a clausal structure. "<para></para>" denotes paragraph opening and ending.

[6] By "Structural elements" we mean documentary information, such as Publishing Date, Source, Link, Editor, etc. Although these are parts of a news text, they are not the article proper, on which we annotate rhetorical relations.

[7] After a comparative study on the rhetorical structure of news published by some Hong Kong newspapers in both English and Chinese, Scollon and Scollon (1997) observed that "quotation is at best ambiguous in Chinese. No standard practice has been observed across newspapers in this set and even within a newspaper, it is not obvious which portions of the text are attributed to whom." We notice that Mainland newspapers have a similar phenomenon.

Besides, 53.9% of the marks appear in structural elements that we didn't intend to analyze[8].

Finally, we decided to use Period, the End-of-line symbol, and these six marks (Question Mark, Exclamation Mark, Colon, Semicolon, Ellipsis and Dash) as delimiters of our EUDA. Quotation mark, Parentheses, and Brackets were not selected.

A special program was designed to conduct the segmentation after each delimiter, with proper adjustment in cases when the delimiter is immediately followed by a right parenthesis, a right quotation mark, or another delimiter.

A pseudo-relation, SAME-UNIT, has been used during annotation to re-tie any discourse segment cut by the segmentation program into fragments.

2.3 Annotation and Validity Control

We use O'Donnell's RSTTool V3.43[9] as our annotation software. We started from the Extended-RST relation set embedded in the software, adding gradually some new relations, and finally got an inventory of 47 relations. We take the same rhetorical predicate with switched arguments as different relations, for instance, SOLUTIONHOOD-S, SOLUTIONHOOD-M and SOLUTIONHOOD-N are regarded as 3 relations.

Following Carlson et al. (2001) and Marcu's (1999) examples, we've composed a 60-page Chinese RST annotation manual, which includes preprocessing procedures, segmentation rules, definitions and examples of the relations, tag definitions for structural elements, tagging conventions for special structures, and a relation selection protocol. When annotating, we choose the most indicative relation according to the manual. Trees are constructed with binary branches except for multinuclear relations.

One experienced annotator had sketched trees for all the 395 files before the completion of the manual. Then she annotated 97 shortest files from 197 randomly selected texts, working independently and with constant reference to the manual. After a one-month break, she re-annotated the 97 files, with reference to the manual and with occasional consultation with Chinese journalists and linguists. The last version, though far from error-free, is currently taken as *the right* version for reliability tests and other statistics.

An intra-coder accuracy test has bee taken between the 1st and 2nd versions of 97 finished trees. The intra-coder accuracy rate (R_v) for a particular variable is defined as

$$R_v = \frac{2*(AT-AS)}{TT-TS} *100\%$$

Where

AT= number of agreed tags;

TT= number of total tags;

TS= number of total tags for structural elements;

AS= number of agreed tags for structural elements.

R_r for relation tags is 84.39%, R_u for unit tags is 85.61%, and R_n for nuclearity tags is 88.12%.

Because SPSS can only calculate Kappa Coefficient for symmetric data, we've only measured Kappa for relation tags to the EUDAs. The outcome, K_r=.738, is quite high.

3 Results

The 97 double-annotated files have in the main body of their texts a total of 677 paragraphs and 1,914 EUDAs. Relational patterns of those PMs are reported in Table 2-7 below[10]. The "N", "S" or "M" tags after each relation indicate the nuclearity status of each EUDA ended with a certain PM. The number of those PMs used in structural elements of CJPL texts are also reported as they make up the total percentage.

Relation (C-？)	P(r|pm)	P(pm|r)
Antithesis-N	1.14%	2.70%
Background-N	2.27%	3.39%
Concession-N	7.95%	7.29%
Conjunction-M	30.68%	5.24%
Disjunction-M	4.55%	36.36%
Elaboration-N	2.27%	1.10%
Elaboration-S	2.27%	1.10%
Evaluation-N	1.14%	0.72%
Interpretation-N	1.14%	0.67%
Joint-M	4.55%	6.90%
Justify-N	4.55%	1.75%
Justify-S	4.55%	1.75%
Nonvolitional-cause-S	2.27%	1.43%
Nonvolitional-result-S	1.14%	0.71%
Otherwise-S	1.14%	16.67%
Solutionhood-M	4.55%	5.33%
Solutionhood-S	14.78%	17.33%
Volitional-cause-N	1.14%	1.32%
Structural elements	7.96%	0.99%
TTL	**100.00%**	**N/A**

Table 2: Rhetorical pattern of C-Question

[8] Parentheses, and other PMs used in structural elements of CJPL texts, are of high relevance to discourse parsing, since they can be used in a preprocessor to filter out text fragments that do not need be annotated in terms of RST.

[9] Publicly downloadable at www.wagsoft.com.

[10] Based on data from the 2nd version of annotated texts.

Relation (C-！)	P(r|pm)	P(pm|r)
Addition-S	5.26%	14.29%
Conjunction-M	15.79%	0.58%
Elaboration-S	5.26%	0.55%
Evaluation-S	10.53%	1.44%
Evidence-S	10.53%	2.33%
Joint-M	5.26%	1.72%
Justify-N	5.26%	0.44%
Justify-S	5.26%	0.44%
Nonvolitional-cause-N	5.26%	0.71%
Solutionhood-N	5.26%	1.33%
Volitional-cause-S	5.26%	1.32%
Structural elements	21.05%	0.57%
TTL	**100.00%**	**N/A**

Table 3: Rhetorical pattern of C-Exclamation

Relation (C-：)	P(r|pm)	P(pm|r)
Attribution-S	10.93%	68.00%
Background-N	0.64%	3.39%
Background-S	0.32%	1.69%
Concession-N	0.32%	1.04%
Elaboration-N	18.97%	32.42%
Evaluation-N	0.64%	1.44%
Justify-S	0.32%	0.44%
Nonvolitional-cause-N	0.32%	0.71%
Preparation-S	4.18%	13.40%
Same-unit-S	0.32%	4.35%
Volitional-cause-N	0.32%	1.32%
Structural elements	62.70%[11]	27.70%
TTL	**100.00%**	**N/A**

Table 4: Rhetorical pattern of C-Colon

Relation (C-；)	P(r|pm)	P(pm|r)
Antithesis-S	1.00%	2.70%
Background-N	1.00%	1.69%
Background-S	1.00%	1.69%
Conjunction-M	59.00%	11.46%
Contrast-M	7.00%	7.69%
Disjunction-M	2.00%	18.18%
List-M	23.00%	24.73%
Purpose-N	1.00%	6.67%
Same-unit-M	2.00%	8.70%
Sequence-M	3.00%	6.12%
TTL	**100.00%**	**N/A**

Table 5: Rhetorical pattern of C-Semicolon

Relation (C-……)	P(r|pm)	P(pm|r)
Conjunction-M	12.50%	0.19%
Disjunction-M	12.50%	9.09%
Elaboration-S	25.00%	1.10%
Evidence-S	25.00%	2.33%
Evaluation-N	12.50%	0.72%
Volitional-result-S	12.50%	1.32%
TTL	**100.00%**	**N/A**

Table 6: Rhetorical pattern of C-Ellipses

Relation (C-——)	P(r|pm)	P(pm|r)
Elaboration-N	32.00%	4.40%
Elaboration-S	4.00%	0.55%
Evaluation-N	12.00%	2.16%
Evaluation-S	4.00%	0.72%
Nonvolitional-cause-S	4.00%	0.71%
Nonvolitional-result-S	4.00%	0.71%
Otherwise-S	4.00%	16.67%
Preparation-N	4.00%	1.03%
Purpose-N	4.00%	6.67%
Restatement-N	4.00%	14.29%
Same-unit-M	24.00%	26.09%
TTL	**100.00%**	**N/A**

Table 7: Rhetorical pattern of C-Dash

The above data suggest at least the following:

1) There is no one-to-one mapping between any of PM studied and a rhetorical relation. But some PMs have dominant rhetorical usages.
2) C-Question Mark is not most frequently related with SOLUTIONHOOD, but with CONJUNCTION. That is because a high percentage of questions in our corpus are rhetorical and used in groups to achieve certain argumentative force.
3) C-Colon is most frequently related with ATTRIBUTION and ELABORATION, apart from its usage in structural elements.
4) C-Semicolon is overwhelmingly associated with multinuclear relations, particularly with CONJUNCTION.
5) C-Dash usually indicates an ELABORATION relation. But since it is often used in pairs, it is often bound to both the Nucleus and Satellite units of a relation.
6) 82.3% tokens of the six Chinese PMs are uniquely related to EUDAs of certain nucleus status in a rhetorical relation, taking even C-Dash into account.
7) The following relations have more than 10% of their instances related to one of the six PMs studied here: ADDITION, ATTRIBUTION, CONJUNCTION, DISJUNCTION, ELABORATION, LIST, OTHERWISE, PREPARTION, RESTATEMENT and SOLUTIONHOOD.
8) Chinese PMs are used somewhat differently from their German equivalents, Exclamation Mark for instance (Fig.2):

[11] This is higher than the overall 42.93% rate for colons used in structural elements, for we've only finished 97 shortest ones from the 197 randomly selected files.

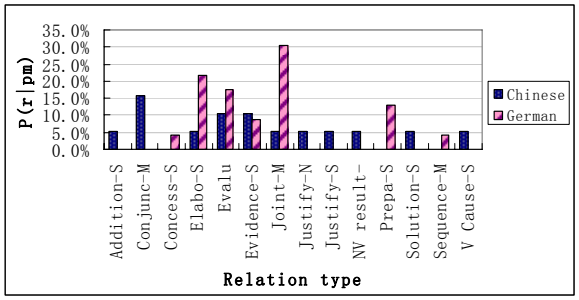

Figure 2: Rhetorical Function of Exclamation Mark in Chinese and German corpora

4 Discussion

How useful are these six PMs in the prediction of rhetorical relations in Chinese texts? In our opinion, this question can be answered partly through a comparison with Chinese cue phrases.

Cue phrases are widely discussed and exploited in the literature of both Chinese studies and RST applications as a major surface device. Unfortunately, Chinese cue phrases in natural texts are difficulty to identify automatically. As known, Chinese words are made up of 1, 2, or more characters, but there is no explicit word delimiter between any pair of adjacent words in a string of characters. Thus, they are not known before tokenization ("*fenci*" in Chinese, meaning "separating into words", or "word segmentation" so as to recognize meaningful words out of possible overlaps or combinations). The task may sound simple, but has been the focus of considerable research efforts (e.g. Webster and Kit, 1992; Guo 1997; Wu, 2003).

Since many cue phrases are made up of high-frequency characters (e.g. "而-*ER*" in "而-*er*" meaning "but/so/and", "然而-*ran'er*" meaning "but/however", "因而-*yin'er*" meaning "so/because of this", "而且-*erqie*" meaing "in addition" etc.; "此-*ci*" in "此后-*cihou*" meaning "later/hereafter", "因此-*yinci*" meaning "as a result", "由此看来-*youcikanlai*" meaning "on this ground/hence", etc.), a considerable amount of computation must be done before these cue phrases can ever been exploited.

Apart from tokenization, POS and WSD are other necessary steps that should be taken before making use of some common cue phrases. They are all hard nuts in Chinese language engineering. Interestingly, many researches done in these three areas have made use of the information carried by PMs (e.g. Sun et al. 1998).

Chan et al. (2000) did a study on identify Chinese connectives as signals of rhetorical relations for their Chinese summarizer. Their tests were successful. But like PMs, Chinese cue phrases are not in a one-to-one mapping relationship with rhetorical relations, either.

In our finished portion of CJPL corpus, we've identified 161 Types of cue phrases[12] at or above our EUDA level, recording 539 tokens. These cue phrases are scattered in 477 EDUAs, indicating 20.5% of the total relations in our finished portion of the corpus. Our six PMs, on the other hand, have 551 tokens in the same finished portion, delimiting 345 EUDAs (and 206 structural elements), and indicating 14.8% of the total relations. However, since there are far more types of cue phrases than types of punctuation marks, 90.1% of cue phrases are sparser at or above our EDUA level than the least frequently used PM—Ellipsis in this case.

And Chinese cue phrases don't signal all the rhetorical relations at all levels. For instance, CONJUNTION is the most frequently used relation in our annotated text (taking 22.1% of all the discursive relations), but it doesn't have strong correlation with any lexical item. Its most frequent lexical cue is "也-*ye*", taking 2.4%. ELABORATION is another common relation in CJPL, but it is rarely marked by cue phrases. ATTRIBUTION, SOLUTIONHOOD and DISJUNCTION are amongst other lowest marked relations in Chinese—they happen to be signaled quite significantly by a punctuation mark.

Given the cost to recognize Chinese cue phrases accurately, the sparseness of many of these cues, and the risk of missing all cue phrases for a particular discursive relation, punctuation marks with strong rhetorical preferences appear to be useful supplements to cue phrases.

5 Conclusion

Because rhetorical structure in Chinese texts is not explicit by itself, systematic and quantitative evaluation of various factors that can contribute to the automatic analysis of texts is quite necessary. The purpose of this study is to look into the discursive patterns of Chinese PMs, to see if they can facilitate discourse parsing without deep semantic analysis.

We have in this study observed the discursive usage of six Chinese PMs, from their overall distribution in our Chinese discourse corpus, their syntax in context, to their rhetorical roles at

[12] We are yet to give a theoretical definition of Cue Phrases in our study. But the identified ones range similarly to those English cue phrases listed in Marcu (1997).

or above our EUDA level. Current statistics seem to suggest clear patterns of their rhetorical roles, and their distinctive correlation with nuclearity in most relations. These patterns and correlation may be useful in NLP projects.

6 Future Work

We are conscious of the size and granularity of our treebank on which this analysis is based. We plan to get a larger team to work on the project, so as to make it more comparable to the English and German RST treebanks.

Since the distinctive nucleus status of EUDAs ended with these PMs may be useful in deciding growth point for RS-tree construction or for tree pruning in summarization, we are also interested in testing how well a baseline relation classifier performs if it always predicts the most frequent relations for these PMs.

Acknowledgement

Special thanks to Dr. Manfred Stede for licensing us to use the Potsdam Commentary Corpus. And thanks to Dr. Michael O'Donnell, FAN Taizhi, HU Fengguo, JIN Narisong, and MA Guangbin for their technical support. The author also fully appreciates the anonymous reviewers for their constructive comments.

References

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual, *Technical Report ISI/TR-545*. www.isi.edu/~marcu.

Lynn Carlson, Daniel Marcu, and Mary. E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers. www.isi.edu/~marcu.

Samuel W. K. Chan, Tom B. Y. Lai, W. J. Gao and B. K. T'sou. 2000. Mining discourse markers for Chinese Textual Summarization. *Workshop on Automatic Summarization*, ACL 2000.

Thomas C. Chuang and Kevin C. Yeh. 2005. Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria. *Computational Linguistics and Chinese Language Processing*. Vol. 10, No. 1, March 2005, pp. 95-122.

Simon H. Corston-Oliver. 1998. Computing Representation of the Structure of Written Discourse. Technical Report. MSR-TR-98-15.

Robert Dale. 1991. The role of punctuation in discourse structure. *Working Notes for the AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*. P13-13. Asilomar.

Jin GUO. 1997. Critical Tokenization and its Properties. *Computational Linguistics*, 23(4): 569-596.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1997. *The rhetorical parsing, summarization, and generation of natural language texts*. PhD thesis. University of Toronto. December 1997. www.isi.edu/~marcu

Daniel Marcu. 1999. *Instructions for manually annotating the discourse structures of texts*. www.isi.edu/~marcu

David Reitter. 2003. *Rhetorical Analysis with Rich-Feature Support Vector Models*. University of Potsdam, Diploma thesis in computational linguistics.

Bilge Say. 1998. *An Information-Based Approach to Punctuation*. Ph.D. dissertation, Bilkent University, Ankara, Turkey. http://www.cs.bilkent.edu.tr/~say/bilge.html.

Ron Scollon and Suzanne Wong Scollon. 1997. Point of view and citation: Fourteen Chinese and English versions of the 'same' news story. *Text*, 17 (1), 83-125.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *Proceedings of the ACL 2004 Workshop 'Discourse Annotation'*. Barcelona.

SUN Maosong, Dayang SHEN, and Benjamin K. Tsou, 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of COLING-ACL'98*.

Benjamin K.Tsou, Weijun Gao, T.V.Y Lai and S.W.K. Chan. 1999. Applying machine learning to identify Chinese discourse markers. *Proceedings of 1999 International Conference on Information Intelligence and Systems*. p 548-53, 31 Oct.-3 Nov. 1999 , Bethesda, MD, USA.

Jonathan J. Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *Proceedings of the 14th International Conference on Computational Linguistics* (COLING'92), pages 1,106-1,110, Nantes, France.

WU Andi. 2003. Chinese Word Segmentation in MSR-NLP. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

ZHANG Yimin, LU Ru-Zhan and SHEN Li-Bin. 2000. A hybrid method for automatic Chinese discourse structure analysis. *Journal of Software*, v 11, n 11, Nov. 2000, p 1527-33.