

# Chinese-DiMLex: A Lexicon of Chinese Discourse Connectives

Shujun Wan<sup>1,2</sup>, Peter Bourgonje<sup>3</sup>, Hongling Xiao<sup>4</sup> and Clara Wan Ching Ho<sup>5</sup>

<sup>1</sup>Shanghai International Studies University, Shanghai, China.

<sup>2</sup>Humboldt-Universität zu Berlin, Berlin, Germany.

<sup>3</sup>University of Potsdam, Potsdam, Germany.

<sup>4</sup>Xiamen University, Xiamen, China.

<sup>5</sup>Saarland University, Saarbrücken, Germany.

Contributing authors: [shujun.wan@shisu.edu.cn](mailto:shujun.wan@shisu.edu.cn);  
[bourgonje@uni-potsdam.de](mailto:bourgonje@uni-potsdam.de); [xhling@xmu.edu.cn](mailto:xhling@xmu.edu.cn);  
[wanching@coli.uni-saarland.de](mailto:wanching@coli.uni-saarland.de);

## Abstract

Machine-readable inventories of connectives that provide information on multiple levels are a useful resource for automated discourse parsing, machine translation, text summarization and argumentation mining, etc. Despite Chinese being one of the world's most widely spoken languages and having a wealth of annotated corpora, such a lexicon for Chinese still remains absent. In contrast, lexicons for many other languages have long been established. In this paper, we present 226 Chinese discourse connectives, augmented with morphological variations, syntactic (part-of-speech) and semantic (PDBT3.0 sense inventory) information, usage examples and English translations. The resulting lexicon, Chinese-DiMLex, is made publicly available in XML format, and is included in *connective-lex.info*, a platform specifically designed for human-friendly browsing of connective lexicons across languages. We describe the creation process of the lexicon, and discuss several Chinese-specific considerations and issues arising and discussed in the process. By demonstrating the process, we hope not only to contribute to research and educational purposes, but also to inspire researchers to use our method as a reference for building lexicons for their (native) language(s).

**Keywords:** Chinese Lexicon, Machine-readable Resource, Discourse Connectives, Discourse Parsing

# 1 Introduction

Using discourse connectives<sup>1</sup> (Pasch, Brauße, Breindl, & Waßner, 2003; Stede, 2002) is one of the several possible ways of creating coherence in a text. As explicit cohesive devices, connectives can signal various semantic relations such as *Causal* (e.g. *because, therefore*), *Temporal* (e.g. *and then, subsequently*), and other types of relations<sup>2</sup>.

In simple terms, connectives are words or short phrases that explicitly express a discourse relation between propositions (usually sentences, see Asher (1993) and its notion of “abstract objects” for further information). Connectives form a syntactically heterogeneous group (e.g., conjunctions and adverbs) and can be semantically ambiguous, as they may exhibit a *discourse* reading or a *sentential* reading. An example of the connective *and* with discourse reading would be “I went home and took a shower.” (expressing a *Temporal/ Con-junctive* relation), whereas the word *and* has sentential reading, simply joining two non-propositional NPs, in “My friends and I went home.”. In addition to this type of ambiguity, connectives can express more than one discourse sense. For example, in the sentence “Once there, what ways of escape are open to them other than drink, drugs or insanity”, the connective *once* has a *Temporal* reading, whereas in the sentence “Normally, once the underlying investment is suspended from trading, the options on those investments also don’t trade”<sup>3</sup>, it has a *Conditional* reading.

Lists or inventories of connectives for a given language that provide information about these types of ambiguity, the discourse sense(s) they can convey, and additional syntactic (part-of-speech) details can enhance human understanding of how the connectives are used. And if such lists are machine-readable, e.g. in XML format, they can be a useful resource for discourse parsing (Bourgonje & Stede, 2020; Hernault, Prendinger, duVerle, & Ishizuka, 2010; Lin, Ng, & Kan, 2014) or for other (downstream) applications such as machine translation (Meyer, Popescu-Belis, Zufferey, & Cartoni, 2011), text summarization (Alemany, 2005) or argumentation mining (Kirschner, Eckle-Kohler, & Gurevych, 2015). While several connective lexicons exist for different languages (see also Section 2), to the best of our knowledge, there is currently no such lexicon for Chinese, despite it being one of the most spoken languages in the world and having a wealth of annotated corpora.

The main purpose of this paper is to present a lexicon of Chinese discourse connectives, explain the procedures to arrive at this lexicon and discuss the decisions made along the way. The resulting lexicon is made available through an easy-to-use web interface that allows browsing and searching the lexicon, as well as comparing Chinese connectives to connectives from other languages, using syntactic (part-of-speech) and semantic (relation sense) properties.

---

<sup>1</sup>Other terms: *discourse markers* (Schiffrin, 1987), *discourse operators* (Redeker, 1991), *clue words* (Cohen, 1984), *cue phrases* (Grosz & Sidner, 1986; Knott & Dale, 1994), etc.

<sup>2</sup>See for example Sanders, Spooren, and Noordman (1992) and Prasad et al. (2008) for frameworks discussing inventories of relations types.

<sup>3</sup>This and the previous example sentence are taken from Prasad et al. (2008).

The rest of this paper is structured as follows: Section 2 describes related work on connective lexicons for other languages. Section 3 discusses some considerations specifically relevant to Chinese discourse connectives. Section 4 explains the dataset used for compiling our initial list of connectives. Section 5 provides details on the methods used to further refine and complete this initial list. Section 6 briefly explains how and where our lexicon is made available. Finally, Section 7 sums up the key contributions of this paper.

## 2 Related Work

### 2.1 Connective Lexicons for other languages

While inventories of discourse connectives have a longer history, the first resource to comprehensively list connectives in a specific language and present them in both human- and machine-readable formats is indeed the German DiMLex (Stede, 2002). This lexicon has been developed and maintained over the course of the last two decades and in its most recent version (Scheffler & Stede, 2016) contains 275 canonical entries (excluding orthographic variants), providing syntactic and semantic information, usage examples, and other details.

An overview of connective lexicons for other languages, along with their size, is presented in Table 1. Additional inventories of connectives that have not made their way into a machine- and human-readable lexicon are presented by Maschler (2002) and Briz, Bordería, and Portolés (2008), for Hebrew and Spanish, respectively.

| Language        | Lexicon       | Entries | Publication                 |
|-----------------|---------------|---------|-----------------------------|
| Arabic          | Arabic        | 390     | Keskes et al. (2014)        |
| Bangla          | DiMLex-Bangla | 123     | Das et al. (2020)           |
| Czech           | CzedLex       | 205     | Mírovský et al. (2016)      |
| DisCoDict       | Dutch         | 207     | Bourgonje et al. (2018)     |
| English         | DiMLex-Eng    | 149     | Das et al. (2018)           |
| French          | LexConn       | 328     | Roze et al. (2012)          |
| Italian         | LICo          | 173     | Feltracco et al. (2016)     |
| Nigerian Pidgin | NaijaLex      | 57      | Marchal et al. (2021)       |
| Portuguese      | LDM-PT        | 210     | Mendes and Lejeune (2016)   |
| Turkish         | TCL           | 113     | Zeyrek and Başibüyük (2019) |
| Ukrainian       | UK-DiMLex     | 163     | Scheffler et al. (2021)     |

**Table 1** Connective Lexicons for various languages.

### 2.2 Corpora and Treebanks for Chinese

Compared to the above-mentioned languages, there is currently no such a lexicon for Chinese connectives that contains multi-layer information. However, there are available corpora and treebanks for Chinese that are systematically

| Type  | Explicit | Implicit | AltLex | Total |
|-------|----------|----------|--------|-------|
| Count | 1223     | 4193     | 118    | 5534  |

**Table 2** Distribution of the relations annotated in the CDTB (Y. Zhou & Xue, 2015, p. 422). CDTB mainly focuses on the annotation of inter-sentential relations.

annotated with information such as discourse connectives, explicit or implicit relations and argument scope.

The early work by Xing (1985), for example, focused on semantic categorization and defined 10 semantic relations that can be expressed by Chinese connectives: 并列 (*Additive*), 连贯 (*Sequential*), 目的 (*Purpose*), 因果 (*Cause-effect*), 递进 (*Progression*), 选择 (*Alteration*), 让步 (*Concession*), 转折 (*Contrast*), 条件 (*Condition*) and 假设 (*Hypothesis*). Based on that, Kuo (1994) listed a total of 239 Chinese words with a linking function and classified them according to a semantic taxonomy.

Regarding existing corpus resources, it's worth noting that the Penn Chinese Treebank (Xue, Xia, Chiou, & Palmer, 2005), with over 200,000 words annotated with part-of-speech tags and syntactic brackets in their first two installments, is a seminal contribution. The current version of the corpus, CTB9.0 (Xue et al., 2016) comprises approximately 2 million words and is meticulously annotated with part-of-speech tags and syntactic brackets. Y. Zhou and Xue (2012) annotated discourse relations of 98 files from the Penn Chinese Treebank (Xue et al., 2005) using the PDTB annotation guidelines (Webber, Prasad, Lee, & Joshi, 2019) with adaptations based on Chinese characteristics. Connectives were annotated with arguments (*arg1* and *arg2*) and information on explicitness or implicitness of the relation. The Discourse Treebank for Chinese (L. Zhou, Li, Wei, & Wong, 2014) annotated all 890 documents of the Penn Chinese Treebank (Xue et al., 2005) following the PDTB2.0 framework (Prasad et al., 2008), with explicit intra-sentential discourse connectives.

Other treebanks with discourse relations annotated using the PDTB framework include the Chinese Discourse Treebank (CDTB) (Y. Zhou & Xue, 2015) and the TED-CDB (CTED) (Long, Cai, Reid, Webber, & Xiong, 2020). The CDTB was annotated according to the PDTB2.0 guidelines on written text of news reports, with 5,534 relations annotated. The CTED was annotated according to PDTB3.0 guidelines on Chinese TED talks, with 15,540 relations annotated, and currently is the largest PDTB-style Chinese discourse corpus in the spoken domain. The CDTB primarily focuses on inter-sentential relations, whereas the CTED covers both inter- and intra-sentential relations. Tables 2 and 3 respectively present the distribution of annotated relations in the two corpora.

## 2.3 Connective-lex.info

The multi-lingual web database *connective-lex.info* was designed to provide search functionality that allows users to browse related connectives both within

| Type           | Explicit | Implicit | AltLex | EntRel | Hypophora | Norel | Total |
|----------------|----------|----------|--------|--------|-----------|-------|-------|
| Intra-Sentence | 5301     | 3209     | 943    | 390    | 4         | 0     | 9847  |
| Inter-Sentence | 230      | 3806     | 91     | 614    | 355       | 597   | 5693  |
| Total          | 5531     | 7015     | 1034   | 1004   | 359       | 597   | 15540 |

**Table 3** Distribution of relations annotated in the CTED (Long, Webber, & Xiong, 2020, p. 2798).

and across languages (Stede, Scheffler, & Mendes, 2019). Currently, in addition to our contribution for Chinese, this browser-based GUI hosts lexicons for 12 other languages. Many of them are based on the data provided by the connective lexicons mentioned in Section 2.1. *connective-lex.info* provides multi-dimensional information on a connective, i.e., syntactic information, semantic relation(s), orthographic variations, examples, and English equivalents, and it allows for a convenient and multi-lingual search of the connective inventories based on string values or syntactic (part-of-speech) and/or semantic (relation sense) attributes.

The part-of-speech tags used on *connective-lex.info* include coordinating conjunctions (*cco*), subordinating conjunctions (*csu*), adverbs or adverbials (*adv*), prepositions (*prep*) and other word classes (*other*). The information concerning relation sense is grounded in the PDTB3.0 relation hierarchy (Webber et al., 2019). *Connective-lex.info* is freely accessible from <http://connective-lex.info/>.

### 3 Definition & Characteristics of Chinese Connectives

According to Pasch et al. (2003), a lexical item **X** is a *connective*, when:

- **X** is not inflectable,
- **X** does not assign case to its syntactic environment,
- **X** expresses some specific, two-place semantic relation,
- the arguments of the relational meaning of **X** are propositional structures,
- the verbalizations of the arguments of the relational meaning of **X** can be clauses.

However, having observed some discrepancies between the usage of connectives in Chinese in comparison to other languages, there is still a need to further specify the general definition based on the language-specific features in the Chinese context. The Chinese term for connectives is “关联词”, which can be literally translated as “relation-word” in English. This group of expressions functions to connect discourse fragments, which can be words, phrases, clauses, sentences and paragraphs (Kaden, 1995).

In terms of syntax, beyond the typical connective elements like subordinating conjunctions, coordinating conjunctions, and adverbs, previous research and corpora have also classified verbs (e.g., “使” meaning “to make”) and certain alternative lexicalizations (AltLex) as connectives (Kuo, 1994; Y. Zhou & Xue, 2015). AltLex refers to some clause or phrase that has the function of signaling a discourse relation, but is not fossilized and displays morphological

variation (e.g. 总而言之 (in summary)). Furthermore, localizers, which are a class of words attached to clauses in denoting *Temporal* or *Spatial* relations, can also function as connectives in Chinese discourse. They are often paired with prepositions as dual markers when introducing subordinate clauses. For instance, the localizer “时” (at...time) is often paired with the preposition “当” (when) to form a connecting pair “当...时” (when), which indicates a *Temporal* relation, as illustrated in example 1.

- (1) 当他们发现被洪水围困时，已经无法撤离了。

When they realized they were surrounded by floodwaters, there was already no way to evacuate.

[csu], *TEMPORAL.Synchronous*

Another characteristic of Chinese connectives lies in the more common use of paired expressions. Subordinating conjunctions in Chinese are observed to be realized lexically by both single and paired expressions (Y. Zhou & Xue, 2015). Instead of having only one component expressing the discourse relation of two arguments, they are often found in pairs, commonly preceding the arguments. Single components can be paired up to mark a discourse relation. We refer to this as “double marking” in this article, a term employed by Xiao (2020). For instance, 因为...所以... (because...so...) is a common double marking pair denoting a *Causal* relation, where the clause attached to 因为 (because) is the cause and the other component, attached to 所以 (so), is the consequence.

Building upon the general definition of a connective and the characteristics of Chinese connectives, in this study, we classify a lexical item as a connective if it fulfills the general conditions proposed in the definition of Pasch et al. (2003) and can be used as single and/or paired expressions (including connecting pairs with localizers, e.g. 当...时 (when...)). Additionally, in line with the criteria defined by DimLex and other lexicons available on *connective-lex.info*, a connective should fall within the part-of-speech tags of subordinating conjunctions, coordinating conjunctions, adverbs, or prepositions<sup>4</sup>. For phrasal items, we require it to be a phrase that is used as idiomatic fixed expressions with no or minimal variations. We do not include verbs in our lexicon, because we focus mainly on discourse connectives that connect two clauses, whereas verbs are predominantly used as predicates connecting a noun phrase and a verb phrase. For the same reason, non-discourse connectives that are used solely to connect nouns/nominalizations (e.g., 和(and)) are excluded as well<sup>5</sup>. More details and examples of excluded cases are presented in Section 4.2.

With regard to the theoretical underpinnings of our lexicon, we follow the PDTB framework and its sense hierarchy to assign relation senses to the connectives in our lexicon. The decision is supported by theoretical reasons, as well as practical concerns. Other frameworks dealing with coherence relations

<sup>4</sup>In DimLex, Stede (2002) drops the second requirement listed in the general definition provided by Pasch et al. (2003) and includes prepositions, which can assign case to their syntactic environment.

<sup>5</sup>Connectives that can connect both sentences and nouns/nominalizations are included in our lexicon. However, the provided examples are limited to their usage in connecting sentences.

((Mann & Thompson, 1988) for example) do not necessarily have a special status for connectives. Among other alternatives Asher and Lascarides (2003); Sanders et al. (1992), the PDTB framework is the most empirically oriented one, as exemplified by the two originating corpora we use, which are also annotated according to this framework. In addition, since we aim to include our lexicon in the *connective-lex.info* platform, which uses the PDTB sense hierarchy throughout, it is justified to adopt this hierarchy as a native inventory for our Chinese lexicon.

## 4 Data

Several strategies are discussed by Bourgonje (2021) for creating connective lexicons from scratch for new languages, or further developing existing lexicons and validating them through bilingual projection (e.g. word alignments). For our Chinese lexicon, we started with a draft inventory that was automatically extracted from the two corpora discussed above: Chinese TED talks (Long, Cai, et al., 2020) and the Chinese Discourse Treebank (Y. Zhou & Xue, 2015). Particularly, we respectively used these two repositories<sup>6</sup> as a starting point.

### 4.1 CDTB and CTED as Data Resource

As mentioned above, the Chinese Discourse Treebank (CDTB) (Y. Zhou & Xue, 2015) and the TED-CDB (CTED) (Long, Webber, & Xiong, 2020) are two of the most recently released available corpora that were systematically annotated with discourse relations and explicit connectives. CDTB and CTED have respectively adopted the annotation framework of PDTB2.0 (Prasad et al., 2008) and PDTB3.0 (Webber et al., 2019), but adaptations were made for both accuracy and practicality reasons to accommodate the characteristics of Chinese connectives.

Specifically, CDTB has made significant adaptations in that it gave up the classic hierarchical classification scheme of the PDTB that annotates each relation with sense labels organized hierarchically in three levels. Instead, a flattened single-level scheme that includes eleven senses, i.e., semantic types, of discourse relations was introduced based on the Chinese data. The eleven sense labels employed in this corpus are: *Alternative*, *Causation*, *Conditional*, *Conjunction*, *Contrast*, *Expansion*, *Purpose*, *Temporal*, *Progression*, *EntRel* (*Entity Relation*), and *NoRel* (*No Relation*). The same sense labels are therefore adopted in the annotation of connectives. Same as in English, the canonical discourse connectives in Chinese are subordinating conjunctions, coordinating conjunctions and discourse adverbials. In addition, there are also “localizers” in Chinese that can function as discourse connectives, very often in combination with prepositions. For example, 当[*preposition*]...的时候[*localizer*], equivalent

---

<sup>6</sup><https://github.com/acoli-repo/rdf4discourse/tree/master/discourse-markers/src/cted>  
<https://github.com/acoli-repo/rdf4discourse/tree/master/discourse-markers/src/cdtb>. Accessed June 13, 2024.

to “when”, signals a subordinate *Temporal* relation. CDTB also includes alternative lexicalizations<sup>7</sup> in the PDTB paradigm. For example, 更引人注目的是 (what is even more striking) is used to introduce an argument that provides details for the preceding argument, therefore expressing an *Expansion* relation. In addition, the verb 使(make) and its variants like 致使 and 使得 are also included in the category AltLex as a signal of *Causation* relation. In total, CDTB has annotated 256 explicit connectives, including 132 cases that are used in pairs.

CTED, on the other hand, adopted the hierarchical three-level scheme of PDTB3.0 and the same set of label names were used for the annotation of each relation. A minor modification was done by adding a second-level sense *Progression* under *EXPANSION*, one of the four top-level senses in PDTB. Also, it loosened the constraints on the following three aspects in the process of annotation.

Firstly, noun phrases that are nominalizations of their verbal forms are regarded as arguments. For example, 我们对他自由的限制 (our limitation towards his freedom) is regarded as arg1 in the following example from (Long, Cai, et al., 2020):

- (2) [我们对他自由的限制]<sub>1</sub> made<sub>RESULT</sub> [使他没有办法加入]<sub>2</sub>  
 [our towards him freedom limitation]<sub>1</sub> [made<sub>RESULT</sub> him cannot join]<sub>2</sub>

Secondly, verbs can serve as explicit connectives, such as 使 “make” in the above example, which is annotated as a connective signaling a *Result* relation. Similar verbs like 来 (literally: “come”) and 去 (literally: “go”), which introduce a *Purpose* relation like the English equivalent “in order to” are also annotated.

Thirdly, relations between non-adjacent sentences are annotated. This addition is especially important given that CTED is a corpus of spoken Chinese. In spoken language, speakers often repeat previously uttered content or insert details or explanations between two arguments.

## 4.2 Chinese-DiMLex Specifics

The primary aim of this study is to create a lexicon of canonical discourse connectives in Chinese. We use the connectives annotated in CDTB and CTED as our initial resource. As for the linguistic expressions in the category AltLex, we only include those that have a fixed form in signaling discourse relations, hence also a closed set as the canonical counterparts, for example, 总而言之 (with the variety of 总之) (in conclusion) and 换句话说 (in other words). Accordingly, other cases listed below are excluded from our lexicon.

Particularly, excluded are non-discourse connectives which are used to connect nouns, such as the additive connectives 和 (and), 跟 (and), 同 (with), 与 (and), (以)及 (and). Also excluded are verbs such as 使 (make), 让 (let/make), 来 and 去 (literally meaning “come” and “go” respectively, equivalent to “in order to” when expressing a purpose). Despite the senses such as *Result* and

---

<sup>7</sup> An English example would be “for [this/that/these/those] reason(s)”.

*Purpose*, they are predominantly used as predicates connecting a noun phrase and a verb phrase, rather than two clauses (i.e., arguments). Some nouns, verbs, and short sentences may also have the “function of connection”, but there are many alternatives for the expressions conveyed. For instance, 听到这个 (once hearing it) marks the sequence between the previous action or event and the following utterance. However, this expression is not unique in that there can be other expressions with the same pattern and the same function, such as 想到这个 (once thinking of it), 说到这个 (once speaking of it), 看到这个 (once seeing it) and so on. Expressions of this sort are also excluded from this study, for instance, 明知道 (knowing), 大概的说 (generally speaking), 对了 (it’s correct), 想不到 (unexpectedly), 加上 (plus), 昨天 (yesterday), and 明天 (tomorrow).

## 5 Methodology

To ensure that our final list of Chinese connectives is as comprehensive and reliable as possible, a workflow consisting of five main steps (Figure 1) was developed and served as the cornerstone of the data processing methodology in this work. Simultaneously, in order to reduce the influence of individual subjectivity on the results, three annotators were involved in the process.

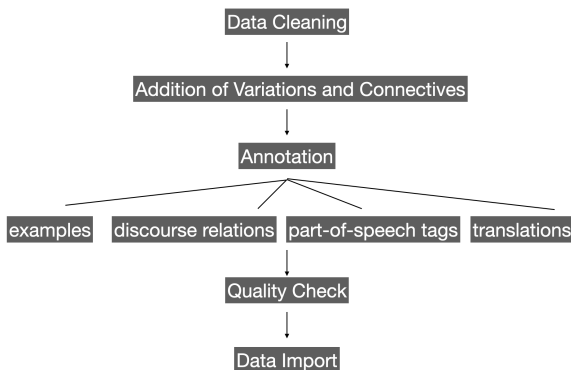


**Figure 1** The methodological workflow of building Chinese-DiMLex

The individual steps can be summarized as follows, and in the subsequent subsections, we will provide a detailed explanation of each component of the workflow. Our first step was to clean up the data coming from the originating data sets. We extracted all the entries from CDTB and CTED and manually excluded the unsuitable cases (Section 5.1). In the second step, we added new entries and lexical variations to existing entries, drawing from the available corpus resources (Section 5.2). In the third step, we augmented the entries with additional information: usage examples (Section 5.3.1), discourse

relations (all possible senses the entry can express) (Section 5.3.2), part-of-speech tags (Section 5.3.3) and translations (English equivalents) (Section 5.3.4). The fourth step consisted of validation, where all information added in the third step was cross-checked among the three native speakers. Controversial or ambiguous cases were thoroughly discussed (Section 5.4). Finally, in the fifth step, the resulting lexicon of Chinese connectives was converted into machine-readable (XML) format, and imported into the *connective-lex.info* platform (Section 5.5).

## 5.1 Data Cleaning

Initially, a total of 637 entries were extracted from CDTB and CTED. However, the raw data contains many cases that need to be sorted or removed, including a) irrelevant items that may have arisen during the data conversion, b) repetitions and c) those cases that do not comply with our definition of a connective as outlined in Section 3 and Section 4.2.

We started the cleaning work with irrelevant information. Entries such as “\*” or “,” were removed. Next, we addressed duplicates (entries present in both originating corpora). In cases where a mixture of Simplified Chinese and Traditional Chinese occurs, we converted the Traditional Chinese characters into Simplified Chinese characters. For example, 由於 (due to) was converted into 由于 (due to). The reason for making this conversion is that, for our work, the difference between the two types of characters is merely orthographical. To ensure stylistic consistency, we chose simplified Chinese, which is used in the majority of the two corpora, as the norm. Finally, we removed those entries mentioned in Section 4.2 that do not match the purpose of our study, i.e. verbs 使 (to make) and 让 (to let), phrases containing modifications such as 远不是...而是... (far not...but...), and numbers as well as ordinal words such as 一 (one), 二 (two), 第一 (first) and 第二 (second).

## 5.2 Morphological Variations and Additional Entries

In this step, we first categorized connectives that are morphologically similar and then determined whether they are variants of each other. If the only difference between a connective A and a connective B is that a part of A is morphologically omitted in B, then A and B are variations of each other. For instance:

- (3) 比如说 / 比如 (for example / for example)
- (4) 但是 / 但 (but / but)
- (5) 自从...以来 / 自...以来 (ever since... / ever since...)
- (6) 仍然 / 仍 (yet / yet)

By following this process, our goal is to improve the simplicity and efficiency of querying the lexicon. However, in certain cases, connectives are treated as separate entries even though parts of them are morphologically and/or semantically identical. For example:

语料字词索引\_语料库在线(www.cncorpus.org)

2023/6/14 18:17:48

查询条件: include c

| 序号  | 词 (/词类) | occurrence   |
|-----|---------|--------------|
| 001 | 和/c     | 79845 8.3553 |
| 002 | 而/c     | 28397 2.9716 |
| 003 | 与/c     | 21385 2.2378 |
| 004 | 但/c     | 15959 1.6700 |
| 005 | 或/c     | 13574 1.4204 |
| 006 | 并/c     | 9259 0.9689  |
| 007 | 因为/c    | 8817 0.9226  |
| 008 | 由于/c    | 7840 0.8204  |
| 009 | 但是/c    | 7673 0.8029  |
| 010 | 则/c     | 7536 0.7886  |
| 011 | 所以/c    | 7512 0.7861  |
| 012 | 因此/c    | 7394 0.7737  |
| 013 | 如果/c    | 7369 0.7711  |
| 014 | 而且/c    | 6116 0.6400  |
| 015 | 以及/c    | 4884 0.5111  |
| 016 | 及/c     | 4408 0.4613  |
| 017 | 虽然/c    | 4059 0.4248  |
| 018 | 只有/c    | 3849 0.4028  |
| 019 | 同时/c    | 3615 0.3783  |
| 020 | 可是/c    | 3355 0.3511  |
| 021 | 因/c     | 3337 0.3492  |
| 022 | 既/c     | 3250 0.3401  |

**Figure 2** A partial screenshot of the list of conjunctions obtained through querying 语料库在线(corpus online), totaling 262 conjunctions. Query performed on June 14, 2023.

(7) 如果...就... | 如果...那么... (if... then... | if... then...)

(8) 尽管...仍然... | 尽管...但是... (although... still... | although... but...)

This is because, unlike the previously mentioned situation, these cases may give rise to slight contextual and modality differences due to the distinct components, despite the connectives as a whole sharing morphological and semantic similarities, such as 就 (just) vs. 那么 (then) in example 7 or 仍然 (still) vs. 但是 (but) in example 8. In specific research areas, these nuances may hold significance. Therefore, when developing the lexicon, we chose not to categorize cases where pragmatic/contextual meanings might vary.

Subsequently, we added connectives that are not presented in CDTB or CTED, but are conforming to our definition. In light of this, we employed 语料库在线(corpus online)<sup>8</sup>, one of the most extensive annotated Chinese corpora that allows us to query word lists, as a foundational reference point. We extracted 262 words from it that had been labeled as “c (conjunctions)”<sup>9</sup> for their part-of-speech. Figure 2 presents a partial screenshot of this list.

Following that, we filtered out irrelevant items, i.e. those previously appearing in CDTB and CTED, inaccurate annotations, and those not aligning with

<sup>8</sup>Corpus online is a web interface with several Chinese language corpora available. Developed by the Chinese State Language Commission. Available at <http://corpus.zhonghuayuwen.org/>. Accessed June 13, 2024.

<sup>9</sup>According to *Jieba*, a model used for Chinese word segmentation. Available at <https://github.com/fxsjy/jieba.git>. Accessed June 13, 2024.

our definitions. Ultimately, this process yielded a total of 92 newly identified connectives. For example:

(9) 万一 (in case):

- 万一被奶奶看见了, 我也得挨批评。

In case my grandmother sees me, I will be criticized as well.

[*csu*], *CONTINGENCY.Condition.Arg2-as-cond*

(10) 越...越... (the more... the more...):

- 他越早知道, 我们就越早甩掉这个案子。

The sooner he knows, the sooner we can get rid of the case.

[*adv*], *EXPANSION.Conjunction*

With this, we arrived at a list<sup>10</sup> of 226 connectives, which were to be annotated for syntactic and semantic information in the next step.

## 5.3 Annotation

To enhance the usability of the resulting lexicon, following the model of many other lexicons described in Section 2.1, we augmented the list of connectives with usage examples, semantic information, syntactic information and English equivalents. The corresponding steps are explained in more detail below.

### 5.3.1 Usage Examples

Adding usage examples to the connectives helps to visualize the contexts in which these connectives are used and provides a solid foundation for other annotation tasks. In order to ensure transparency and breadth of the register the examples cover, we chose “语料库在线 (corpus online)” and “现代汉语语料库 (Corpus of Modern Chinese)”<sup>11</sup> as resources in addition to CDTB and CTED.

In selecting the examples, we tried to cover all possible discourse situations in which each connective can be used (such as Example 11). Meanwhile, for connectives with morphological variations, example sentences are included for both canonical and non-canonical forms.

(11) 至少 (at least):

- 那些青山, 如果没有其他影响, 至少曾令我远离政治, 这已经是其功不小了。

Those green hills, if for nothing else, have at least kept me out of politics, which is already a great contribution.

[*adv*], *COMPARISON.Concession.Arg2-as-denier*

<sup>10</sup>Arriving at an exhaustive list of connectives is a long-term task. We intend to continue our work on Chinese-DiMLex and are particularly open to suggestions for additions

<sup>11</sup>Developed by the Center for Chinese linguistics PKU. Available at [http://ccl.pku.edu.cn:8080/ccl\\_corpus/](http://ccl.pku.edu.cn:8080/ccl_corpus/). Accessed June 13, 2024.

- 不一定大家都爱看，至少我所认识的人中就有不少是不屑于这些无聊东西的。

Not necessarily everyone loves to watch it, at least many people I know are disdainful of these boring things.

[*adv*], *EXPANSION.Instantiation.Arg2-as-instance*

- (12) 如果 / 如果...的话 (if / if):

- 如果没有领导作风上的进步，这样大的成绩是不可能的。

Such great achievements would not have been possible if there were no improvements in leadership.

- 如果用友谊的量杯测定一下的话，他们的友谊算是程度高的。

Their friendship is considered to be of high degree if measured by the scale of friendship.

[*csu*], *CONTINGENCY.Condition.Arg2-as-cond*

Out of multiple potential example sentences, we chose those that were more straightforward (i.e., one single clause acting as an argument) and as general (i.e., non-domain-specific) as possible. For instance, we chose sentence 13 instead of sentence 14 as an usage example for the connective 以免 (so as not to), on the basis of 13 being more concise and syntactically simple.

- (13) 妻子示意我收敛几分，以免邻居愤怒。

My wife signaled me to fix my behavior, so as not to irritate the neighbors.

- (14) 城市中等教育要大力发展职业、技术教育，使中等专业学校有更大的发展，普通高中要适当压缩，但不宜压缩过快、过多，以免发生部分初中生既不能就业又不能升学的情况。

The urban secondary education should vigorously develop the vocational and the technical education, so that those secondary schools can get greater development, and general high schools should be properly tightened up, but not too fast and too much, so as not to happen that some junior high school students can neither be employed nor go on to higher education.

### 5.3.2 Discourse Relation Senses

After gathering usage examples, we proceeded to specify the possible discourse relation senses that each connective can convey. To achieve this, we utilized PDTB 3.0 inventory (Webber et al., 2019). Its well-defined sense set provides a comprehensive framework for describing the relation senses for a lexically grounded discourse annotation. Furthermore, given that PDTB 3.0 is also employed by lexicons on the platform *connective-lex.info* in other languages, its integration here also creates opportunities for cross-linguistic comparative studies.

Annotation at the discourse level is an extremely time- and resource-intensive task due to the inevitable subjectivity, complexity, and ambiguity of

discourse. To ensure a clean annotation work despite these challenges and to achieve a reliable annotation outcome, three annotators were engaged. All of them are native Chinese speakers with expertise in linguistics.

The annotation work for relation senses was divided into two main phases. The first phase consisted of reviewing and unifying the annotations derived from CDTB and CTED, and the second phase focused on annotating the discourse relations for the 92 additional entries.

### a) Connectives from CDTB and CTED

For the connectives extracted from CDTB and CTED, the information pertaining to relation senses has already been provided by the originating resources. However, as described in Section 4, the two treebanks employed different relation sense inventories. Accordingly, our work in this step is to convert the differently defined and annotated senses in the corpora CDTB and CTED into the PDTB 3.0 style. The whole mapping process strictly aligns with the PDTB3.0 guidelines. Examples of the conversions are presented in 15 and 16.

(15) 后来 (later):

- 他先学会沉默, 后来又学会抽烟。

He first learned to be silent, and later learned to smoke.

*TEMPORAL*

→ *TEMPORAL.Asynchronous.Precedence*

(16) 不但...更... (not only... more over...):

- 他的水平和他的风度, 不但令听者折服, 更令自己折服!

His level and his poise, not only convinced the listener, more so convinced himself!

*EXPANSION.Progress.Arg2-as-progr*

→ *EXPANSION.Conjunction*

### b) Additional Connectives

For the new entries, the annotators independently annotated the discourse relations conveyed by each connective based on the usage examples, strictly following the PDTB 3.0 annotation guidelines. From this, we calculated the inter-annotator agreement score using Fleiss' kappa, a statistical measure typically employed for assessing the reliability of agreement among three or more raters. This initial separate annotation resulted in an inter-annotator agreement score of  $\kappa = 0.539$ . Regarding the interpretation of the kappa value, there is no uniform standard, which is indeed also hardly possible, as different annotation tasks have different goals and complications (Wan, 2023). In this study, we use the scale proposed by Landis and Koch (1977). According to this scale, a value of 0.539 means that the annotators reached moderate agreement. Although this score is somewhat distant from the goal of approaching 1, considering that discourse annotation often involves subjectivity and there

is a possibility of multiple annotations being applicable, we believe that our annotations are still effective and reliable.

Subsequently, the annotators compared their annotations. At this stage, they engaged in discussions and careful consideration for every instance where annotations varied. Ultimately, they achieved a unanimous agreement, yielding a thoroughly reviewed annotation.

The resulting relation sense annotations are illustrated in the following figures. Figure 3 displays the degree of ambiguity of all connectives, indicating that most entries express only one sense, while some can express up to 5 different senses (there is one connective potentially expressing 4 senses and one connective potentially expressing 5 senses). Figure 4 displays the actual senses the connectives can express, grouping connectives by their senses on the top level of the PDTB hierarchy.

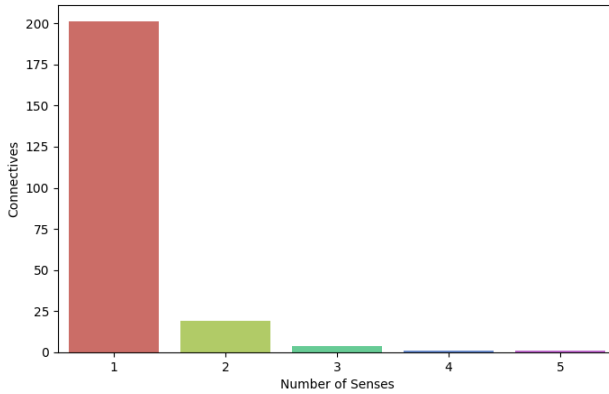


**Figure 3** The degree of ambiguity in the senses of the connectives in Chinese-DiMLex

### 5.3.3 Part-of-speech Tags

In alignment with the other lexicons on *connective-lex.info*, we also assigned part-of-speech tags to all Chinese connectives. Here, we adopted the relatively coarse part-of-speech tags used on *connective-lex.info*. For each entry, we specified one of the following four categories:

- **cco**: coordinating conjunction, e.g. 并且 (and)
- **csu**: subordinating conjunction, e.g. 即使 (even if)
- **adv**: adverb or adverbial, e.g. 至少 (at least)
- **prep**: preposition e.g. 除了 (except for)

For this step, to ensure the reliability of the results, the three annotators followed the same procedure as they did for annotating relation senses. Initially, they individually annotated the part-of-speech tags. For this, the

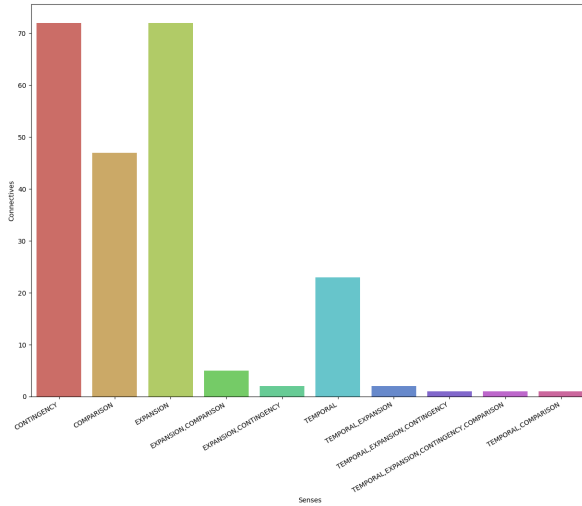


**Figure 4** Sense groups expressed by all connectives in Chinese-DiMLex

inter-annotator agreement was considerably higher than for the sense annotation task, and the three annotators reached “almost perfect agreement” (Landis & Koch, 1977), with  $\kappa = 0.941$  (Fleiss’ kappa).

Following this, the annotators conducted a review and discussion of the points where discrepancies existed. The resulting distribution of part-of-speech tags of the 226 Chinese connectives is presented in Figure 5.

### 5.3.4 English Equivalents

In order to make our lexicon comprehensible for users who do not speak Chinese, we adopted English as a lingua franca for translation and provided English equivalents for all entries. It is worth mentioning that this feature is primarily designed to accommodate the comprehension of non-Chinese-speaking users, and the provided translation is not guaranteed to be accurate in all possible contexts of the connective. We prioritized reflecting the literal features of the original Chinese text, and in some cases, as a result, the English equivalent may not always be well-formed.

- (17) 不但...而且... (not only...but also...)  
       不但...更... (not only...but also... → not only...moreover...)  
       不但...甚至... (not only...but also... → not only...even...)

For instance, the connectives in Example 17 can indeed all be translated as the commonly used *not only... but also...* in English. However, there are subtle differences among the original Chinese entries that the typical English translation could not cover, especially in the second part of the pairs. Whereas

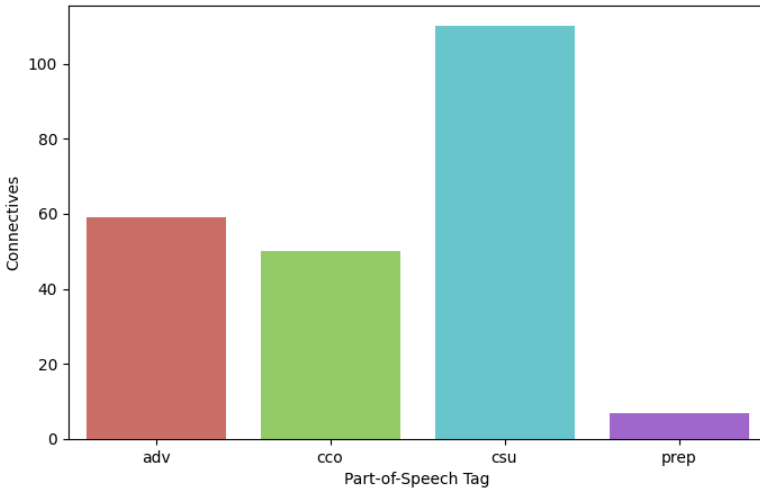


**Figure 5** Distribution of the part-of-speech tags of the 226 Chinese connectives.

而且 conveys a more neutral juxtaposition, 更 implies a degree of progression, and 甚至 expresses an attitude beyond expectation. Therefore, in order to show the differences, we translated the three connectives as *not only...but also*, *not only...moreover...* and *not only...even...* respectively, although the two latter expressions are not typical collocations in English.

In addition, we tried to keep the part-of-speech tags of the English translations the same as that of the Chinese original, but when a translation with a different part-of-speech tags expressed the meaning more accurately, we used that instead. For example, the part-of-speech tag of 由于 is subordinating conjunction, but semantically it is closer to the English prepositional phrase *due to*. In this case, we ignored the syntactic types and chose *due to* as the translation.

## 5.4 Quality Check

As mentioned in Sections 5.3.2 and 5.3.3, we put considerable effort into maintaining the quality of annotations throughout the entire annotation process, despite the inevitable challenges encountered, particularly in annotating relation sense. In order to further enhance the reliability of the augmented information, all annotation tasks underwent a three-step procedure.

First, each annotator annotated the data separately. During this phase, any instances of uncertainty or ambiguity were explicitly marked for subsequent discussion. When annotating relation senses and part-of-speech tags, for which inter-annotator agreement could be assessed, we also calculated the Fleiss' kappa score to measure the level of agreement among the annotators.

The second step was to deliberate on the highlighted cases and achieve consensus on them. For example, the discourse relation of 不论 (regardless of), presented in Example 18, was originally annotated as *Conditional* in the CDTB corpus. The annotators highlighted this as an uncertainty in the first annotation round, because this connective seemed to carry more of a concessive sense than a conditional one. By reviewing the original text of the CDTB and searching for more examples in other corpora, all three annotators agreed that *Concession* was indeed more accurate. Accordingly, the original annotation *Conditional* was changed to *Concession*, which translated to *Comparison.Concession.Arg2-as-denier*, since we used the PDTB3.0 inventory.

(18) 不论 (regardless of):

- 不论什么人，违者必惩。

Regardless of who is involved, violators will be punished.  
(source: corpus online)

- 不论刮风下雨，他从不缺课。

Regardless of it's raining or winding, he is never absent from school.  
(source: CDTB corpus)

In the third and last step, the annotators presented and compared their annotation results once more. In this final check, all annotations (not just ambiguous and uncertain cases) were thoroughly examined and cross-checked.

## 5.5 Data Format

Up to this point, the entries were kept in `csv` format for convenient editing in a spreadsheet editor. Since our goal is to create a lexicon that is both human- and machine-readable, and after completing most of the manual editing steps mentioned earlier, we have transitioned to using the XML format for the final lexicon. We have adopted the structure proposed by DiMLex and utilized a Python script<sup>12</sup> to convert the csv data into XML format. The resulting format is depicted in Figure 6.

Using this format has given us the advantage to easily import the lexicons to the *connective-lex.info* platform, for users to conveniently browse and compare entries to that of other languages, based on semantic or syntactic attributes.

## 6 Lexicon Availability

We made our lexicon publicly available<sup>13</sup> in a machine-readable format as a separate resource, under CC BY 4.0<sup>14</sup> license. For human users, we included it in *connective-lex.info* (Stede et al., 2019), enabling plain text search and retrieval by syntactic and semantic attributes, which in turn enables easy interlingual comparison of connective groups. Figure 7 includes a screenshot of

<sup>12</sup><https://github.com/discourse-lab/chinese-dimlex/blob/main/csv2xml.py>. Accessed June 13, 2024.

<sup>13</sup><https://github.com/discourse-lab/chinese-dimlex>. Accessed June 13, 2024.

<sup>14</sup><https://creativecommons.org/licenses/by/4.0/>

```

<?xml version='1.0' encoding='UTF-8'?>
<dimlex>
  <entry id="c1" word="一...就...">
    <orths>
      <orth type="cont" canonical="1" onr="1o1">
        <part type="single">一...就...</part>
      </orth>
    </orths>
    <syn>
      <cat>csu</cat>
      <sem>
        <ptb3_relation sense="TEMPORAL:Asynchronous:Succession"/>
        <example>我一见到他就讨厌，但不敢则声。</example>
        <english_equivalent>once ... then ...</english_equivalent>
      </sem>
    </syn>
  </entry>

```

**Figure 6** Single lexicon entry in XML format.

the GUI, where our Chinese lexicon is selected in the top-left pane, all coordinating conjunctions are selected in the pane on the left in the middle, and all Comparison top-level PDTB3 sense entries are selected in the bottom-left pane. The right pane shows the found entries, augmented with usage examples and their English equivalents.



**Figure 7** connective-lex.info GUI

## 7 Conclusion

In this paper, we present Chinese-DiMLex, a valuable and reliable resource for both linguistics and NLP research. Chinese-DiMLex comprises 226 Chinese discourse connectives and is enhanced with morphological variations, syntactic (part-of-speech), and semantic (discourse relation sense, following the PDTB3.0 sense inventory) information. It also provides usage examples from various corpora along with English translations. Chinese-DiMLex is made publicly available in XML format and included in a platform specifically designed for human-friendly browsing of connective lexicons, and comparing them across languages. In addition to presenting Chinese-DiMLex, our work also provides insights into the process of creating lexicons of connectives using pre-existing discourse-annotated corpora. We hope that this can contribute to the development of lexicons of connectives for a wider range of languages.

The construction of a lexical resource for connectives in a specific language is both valuable and challenging. Experience with the creation and maintenance of such lexicons for other languages has taught us that it is a long-term process, and that each revision, corpus study and reconsideration of the content is likely to add either new entries or modifications to existing ones. Thus, we encourage researchers working with our resource to reach out and provide feedback wherever appropriate, and see it as an important piece of future work for us to continue the development of our resource in the future.

## 8 Acknowledgments

This work was partially supported by *Chinese Scholarship Council* and *German Research Foundation (DFG) – SFB 1412, 416591334*. We are deeply grateful for this. Additionally, we would like to express our sincere appreciation to Prof. Dr. Manfred Stede for his valuable and constructive suggestions during the development of this research. We are also extremely thankful for the feedback provided by the reviewers, which has significantly contributed to the improvement of our work.

## References

- Alemany, L.A.i. (2005). *Representing Discourse for Automatic Text Summarization via Shallow NLP Techniques* (PhD dissertation). Universitat de Barcelona.
- Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: SLAP 50, Kluwer.
- Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

- Bourgonje, P. (2021). *Shallow Discourse Parsing for German* (PhD dissertation). Universität Potsdam.
- Bourgonje, P., Hoek, J., Evers-Vermeul, J., Redeker, G., Sanders, T., Stede, M. (2018). Constructing a Lexicon of Dutch Discourse Connectives. *Computational Linguistics in the Netherlands Journal*, 8, 163–175.
- Bourgonje, P., & Stede, M. (2020). Exploiting a Lexical Resource for Discourse Connective Disambiguation in German. *Proceedings of the 28th international conference on computational linguistics* (pp. 5737–5748). Barcelona, Spain (Online). Retrieved from <https://www.aclweb.org/anthology/2020.coling-main.505> doi:10.18653/v1/2020.coling-main.505
- Briz, A., Bordería, S.P., Portolés, J. (2008). *Diccionario de partículas discursivas del español*. Retrieved from <http://www.dpde.es>
- Cohen, R. (1984). A Computational Theory of the Function of Clue Words in Argument Understanding. *Proceedings of the 10th international conference on computational linguistics and 22nd annual meeting of the association for computational linguistics* (pp. 251–258). Stanford, California, USA: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P84-1055> doi:10.3115/980491.980546
- Das, D., Scheffler, T., Bourgonje, P., Stede, M. (2018). Constructing a Lexicon of English Discourse Connectives. *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue* (pp. 360–365). Melbourne, Australia: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W18-5042> doi:10.18653/v1/W18-5042
- Das, D., Stede, M., Ghosh, S.S., Chatterjee, L. (2020). DiMLex-Bangla: A Lexicon of Bangla Discourse Connectives. *Proceedings of the 12th international conference on language resources and evaluation (lrec'20)* (pp. 1097–1102). Marseille, France: European Language Resources Association (ELRA). Retrieved from <https://www.aclweb.org/anthology/2020.lrec-1.138>
- Feltracco, A., Jezek, E., Magnini, B., Stede, M. (2016). LICO: A Lexicon of Italian Connectives. *Proceedings of the 3rd italian conference on computational linguistics (clic-it)* (pp. 141–145). Napoli, Italy.
- Grosz, B.J., & Sidner, C.L. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3), 175–204. Retrieved from <https://www.aclweb.org/anthology/J86-3001>

- Hernault, H., Prendinger, H., duVerle, D., Ishizuka, M. (2010). HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3), 1-33.
- Kaden, K. (1995). Eifring Halvor : Clause Combination in Chinese. *Cahiers de Linguistique - Asie Orientale*, 24(2), 295-302. Retrieved from [https://www.persee.fr/doc/clao\\_0153-3320\\_1995\\_num\\_24\\_2\\_1483](https://www.persee.fr/doc/clao_0153-3320_1995_num_24_2_1483)
- Keskes, I., Zitoune, F.B., Belguith, L.H. (2014). Learning Explicit and Implicit Arabic Discourse Relations. *Journal of King Saud University-Computer and Information Sciences*, 26(4), 398-416.
- Kirschner, C., Eckle-Kohler, J., Gurevych, I. (2015, June). Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications. *Proceedings of the 2015 naacl-hlt conference*.
- Knott, A., & Dale, R. (1994). Using Linguistic Phenomena to Motivate a Set of Coherence Relations. *Discourse Processes*, 18, 35-62.
- Kuo, P.-M. (1994). *Discourse Connectives and Discourse Structure in Mandarin Chinese*. University of Illinois at Urbana-Champaign.
- Landis, J.R., & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *biometrics*, 159-174.
- Lin, Z., Ng, H.T., Kan, M.-Y. (2014). A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 20(2), 151-184.  
doi:10.1017/S1351324912000307
- Long, W., Cai, X., Reid, J., Webber, B., Xiong, D. (2020). Shallow Discourse Annotation for Chinese TED Talks. *Proceedings of the 12th language resources and evaluation conference* (pp. 1025-1032). Marseille, France: European Language Resources Association. Retrieved from <https://aclanthology.org/2020.lrec-1.129>
- Long, W., Webber, B., Xiong, D. (2020). TED-CDB: A large-scale Chinese discourse relation dataset on TED talks. *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2793-2803). Association for Computational Linguistics. doi:10.18653/

v1/2020.emnlp-main.223

- Mann, W., & Thompson, S. (1988). Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *TEXT*, 8, 243-281.
- Marchal, M., Scholman, M., Demberg, V. (2021). Semi-automatic Discourse Annotation in a Low-resource Language: Developing a Connective Lexicon for Nigerian Pidgin. *Proceedings of the 2nd workshop on computational approaches to discourse* (pp. 84-94).
- Maschler, Y. (2002). The Role of Discourse Markers in the Construction of Multivocality in Israeli Hebrew Talk in Interaction. *Research on Language and Social Interaction*, 35(1), 1-38.
- doi:10.1207/S15327973RLSI35-1\\_1
- Mendes, A., & Lejeune, P. (2016). LDM-PT. A Portuguese Lexicon of Discourse Markers. *Conference handbook of textlink – structuring discourse in multilingual europe second action conference* (pp. 89-92). Budapest, Hungary.
- Meyer, T., Popescu-Belis, A., Zufferey, S., Cartoni, B. (2011). Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. *Proceedings of the sigdial 2011 conference* (pp. 194-203). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=2132890.2132912>
- Mírovský, J., Jínová, P., Rysová, M., Poláková, L. (2016). Designing CzeDLex – a lexicon of Czech discourse connectives. *Proceedings of the 30th pacific asia conference on language, information and computation: Posters* (pp. 449-457). Seoul, South Korea. Retrieved from <https://www.aclweb.org/anthology/Y16-3017>
- Pasch, R., Brauße, U., Breindl, E., Waßner, U.H. (2003). *Handbuch der Deutschen Konnektoren*. Berlin/New York: Walter de Gruyter.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B. (2008). The Penn Discourse TreeBank 2.0. *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)* (pp. 2961-2968). Marrakech, Morocco: European Language Resources Association (ELRA). Retrieved from [http://www.lrec-conf.org/proceedings/lrec2008/pdf/754\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf)
- Redeker, G. (1991). Linguistic Markers of Discourse structure. *Linguistics*, 26, 1139-1172.

Roze, C., Danlos, L., Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*, 10.

doi:10.4000/discours.8645

Sanders, T.J., Spooren, W.P., Noordman, L.G. (1992). Toward a Taxonomy of Coherence Relations. *Discourse Processes*, 15(1), 1-35.

doi:10.1080/01638539209544800

Scheffler, T., Solopova, V., Zolotarenko, O., Razno, M. (2021). A Computational Lexicon of Ukrainian Discourse Connectives. *International conference on information and communication technologies in education, research, and industrial applications* (pp. 525–537).

Scheffler, T., & Stede, M. (2016). Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. N.C. et al. (Ed.), *Proceedings of the 10th international conference on language resources and evaluation (lrec'16)* (pp. 1008–1013). Portorož, Slovenia: European Language Resources Association (ELRA).

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511611841

Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers. *Exploring the lexicon - theory and computation*. Alessandria: Edizioni dell'Orso.

Stede, M., Scheffler, T., Mendes, A. (2019). Connective-lex: A Web-based Multilingual Lexical Resource for Connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*.

Wan, S. (2023). *Argumentationsstrategien chinesischer deutschlerner/-innen eine korpusbasierte kontrastive untersuchung im vergleich zu deutschen l1-sprecher/-innen*. J.B. Metzler.

Webber, B., Prasad, R., Lee, A., Joshi, A. (2019). The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*, 35, 108.

Xiao, H. (2020). *Subjectivity, Causality and Connectives in Mandarin Chinese: Converging Evidence from Written, Spoken and Social Media Discourse*. Amsterdam: LOT.

- Xing, F. (1985). *Fuju yu Guanxi Yucl* (The Compound Sentence and the Relation Words). *Haerbin: Heilongjiang Renmin Chubanshe*.
- Xue, N., Chang, M., Chiou, F.-D., Xia, F., Palmer, M., Jiang, Z., Zhang, X. (2016). *Chinese Treebank 9.0*. Retrieved from <https://catalog ldc.upenn.edu/LDC2016T13>
- Xue, N., Xia, F., Chiou, F.-D., Palmer, M. (2005). The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2), 207–238.
- doi:10.1017/S135132490400364X
- Zeyrek, D., & Başbüyük, K. (2019). TCL-A Lexicon of Turkish Discourse Connectives. *Proceedings of the first international workshop on designing meaning representations* (pp. 73–81).
- Zhou, L., Li, B., Wei, Z., Wong, K.-F. (2014). The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank. *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 942–949). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from <http://www.lrec-conf.org/proceedings/lrec2014/pdf/270.Paper.pdf>
- Zhou, Y., & Xue, N. (2012). PDTB-style discourse annotation of chinese text. *Proceedings of the 50th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 69–77).
- Zhou, Y., & Xue, N. (2015). The Chinese Discourse TreeBank: A Chinese Corpus Annotated with Discourse Relations. *Language Resources and Evaluation*, 49(2), 397–431.

doi:10.1007/s10579-014-9290-3