Discourse and Lexicons: Lexemes, MWEs, Grammatical Constructions and Compositional Word Combinations to Signal Discourse Relations

Laurence Danlos

Discourse and Lexicons: Lexemes, MWEs, Grammatical Constructions and Compositional Word Combinations to Signal Discourse Relations

Laurence Danlos
Université Paris Diderot, Laboratoire de Linguistique Formelle
2 Place Thomas Mann, 75013 Paris, France
Laurence.Danlos@linguist.univ-paris-diderot.fr

Abstract

Lexicons generally record a list of lexemes or non-compositional multiword expressions. We propose to build lexicons for **compositional** word combinations, namely "secondary discourse connectives". Secondary discourse connectives play the same function as "primary discourse connectives" but the latter are either lexemes or non-compositional multiword expressions. The paper defines primary and secondary connectives, and explains why it is possible to build a lexicon for the compositional ones and how it could be organized. It also puts forward the utility of such a lexicon in discourse annotation and parsing. Finally, it opens the discussion on the constructions that signal a discourse relation between two spans of text.

1 Introduction

Lexicons generally record a list of lexemes — e.g., a list of English verbs in VerbNet (Karin Kipper and Palmer, 2006) —-or a list of multiword expressions (MWEs) — see (Losnegaard et al., 2016) for a survey on MWE ressources. We quote (Savary and Cordeiro, 2018): "Multiword expressions are word combinations, such as *all of a sudden, a hot dog, to pay a visit* or *to pull ones leg*, which exhibit lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasies. (...) A prominent feature of many MWEs (...) is their non-compositional semantics, i.e. the fact that their meaning cannot be deduced from the meanings of their components, and from their syntactic structure, in a way deemed regular for the given language."

We propose to build lexicons for **compositional** word combinations, namely "secondary discourse connectives". They play the same function as "primary discourse connectives" but the latter are lexemes or non-compositional multiword expressions. Primary and secondary connectives are illustrated below in the examples in (1), which all express a causal relation between Fred's jokes and his friends' hilarity.[1] In (1a), the causal relation is explicitly signalled by the primary connective (in magenta) *therefore* which is an adverb; in (1b), it is signalled by the primary connective *as a result,* which is a frozen multiword prepositional phrase. In (1c), it is signalled by *this caused*, a secondary connective (in blue) which is made of an anaphoric subject and a verb, both used with a compositional meaning; in (1d), it is signalled by *because of this*, a prepositional phrase made of a compound preposition followed by an anaphoric pronoun.

(1) a. Fred didn't stop joking. Therefore, his friends enjoyed hilarity throughout the evening.

b. Fred didn't stop joking. As a result, his friends enjoyed hilarity throughout the evening.

c. Fred didn't stop joking. This caused hilarity among his friends for the whole evening.

d. Fred didn't stop joking. Because of this, his friends enjoyed hilarity throughout the evening.

[1] In this paper, either we deliberately use invented examples for purposes of succinctly illustrating particular phenomena, or we use real examples for more complex phemomena (the source of such examples is then given in footnote).

The notion of a (primary) discourse connective was introduced a long time ago and there exist several lexicons in different languages, see Section 2. These lexicons record a list of lexemes or MWEs, as it is the case of other lexicons.

The notion of a secondary connective is more recent: it was originally introduced by (Rysová and Rysová, 2014). It is helpful both in discourse annotation and discourse parsing as explained in Section 4. Although secondary connectives are compositional word combinations, it is possible to build a lexicon for secondary connectives in a given language because they follow a limited number of templates and each template is "lexically headed" by a unique core unit, i.e. the unit that has the strongest meaning, which belongs to a closed list. For example, one of the templates in English is headed by a "discourse verb"[2] such as *cause, provoke, be due to, precede, follow* . . . , which belongs to a closed list. Moreover, a discourse verb is the lexical head of a secondary connective only if its subject is an anaphoric pronoun such as *this, that, it*, as it is the case in (1c). If the subject of a discourse verb is a definite noun phrase, as in (2a), the subject-verb combination is called a "free connecting phrase" in (Rysová and Rysová, 2014).[3] Another template is headed by a (simple or compound) preposition such as *because of*, which is the lexical head of a secondary connective only if it introduces an anaphoric pronoun, as it is the case in (1d). If it introduces a nominal phrase as in (2b), the prepositional phrase (*because of his jokes*) is a free connecting phrase.

(2) a. Fred didn't stop joking. His jokes caused hilarity among his friends for the whole evening.

b. Fred didn't stop joking. Because of his jokes, his friends enjoyed hilarity throughout the evening.

Some templates for secondary connectives include an abstract anaphora — an anaphora which refers to an " abstract entity" (Asher, 1992), such as eventualities or facts — but this is not always the case, see Section 3. One of the basic difference between secondary connectives and free connective phrases is context dependency: the use of free connective phrases is heavily dependant on context, contrary to the use of secondary connectives. As an illustration, consider (3). The causal relation in (3a) is expressed by the free connecting phrase *because of this illness*. Its use makes sense because the left context mentions Fred's pneumonia, which is the antecedent of the anaphoric noun phrase *this illness*. On the other hand, this expression cannot be used in (3b) — which is incoherent (hence the sign #) — because there is no antecedent for the anaphoric noun phrase. It contrasts with the secondary connective *because of this*, which can be used in either context, (3c).

(3) a. Fred has pneumonia. Because of this illness, he will be absent from his work for two weeks.

b. #Fred is on his honeymoon. Because of this illness, he will be absent from his work for two weeks.

c. Fred /has pneumonia/ is on his honeymoon. Because of this, he will be absent from his work for two weeks.

In a nutshell, there exist at least three types of word combinations to signal a discourse relation between two spans of texts: (i) primary connectives, which are one-word expressions or frozen MWEs and for which several lexicons have already been built in different languages, (ii) secondary connectives, which are multi-word combinations used with a compositional meaning but with a unique lexical head, and for which we will describe how to build lexicons and how to use these lexicons in discourse annotation and parsing, and (iii) free connective phrases, which are multi-word combinations used with a compositional meaning and which cannot be recorded in a lexicon because of their productive nature due to (at least) two content words with a compositional meaning: lexical entries like *his jokes caused* ou *because of this*

[2] A discours verb is a verb with two arguments both referring to eventualities, such as *cause* (*An earthquake caused a tsunami*). It plays a role in discourse when one of its argument is an anaphora (*There has been an earthquake. It caused a tsunami in Japan).* The notion of discourse verb was originally introduced in (Danlos, 2006).

[3] This term is not appropriate for a word combination such as *his jokes caused*, which do not form a phrase. However, it is appropriate for *because of his jokes*, so we use it.

illness would be undesirable because of their heavy dependence on context and the absence of a lexical head.

It should be noted that some primary connectives may be considered as grammaticalized forms of secondary connectives. This idea is supported by the historical origin of the present-day primary connectives that arose from similar structures (and parts of speech) as present-day secondary connectives. See English *because* coming from combination of a preposition *bi* and noun *cause* or German *dagegen* containing the preposition *gegen* and a referential part *da*. This means that, due to language changes and possible increasing grammaticalization, secondary connectives may become primary in the future.

The paper is organized as follows. Section 2 gives a brief summary on primary connectives. Section 3 presents the main templates for English secondary connectives. Section 4 shows the usefulness of connective lexicons for both discourse annotation and parsing. Section 5 discusses on two possible ways to organize secondary connective lexicons, one is explored for French, the other one for Czeck. Finally, Section 6 opens the discussion on grammatical constructions with a discourse use.

2 Primary connectives

We define a primary discourse connective as an element which is, morpho-syntactically, a single-word unit or a frozen multiword unit, and semantically, a predicate with two arguments referring to eventualities. This characterization of primary discourse connective corresponds to the traditional notion of "connective" as used in the literature; see, e.g., (Zwicky, 1985; Hrbáček, 1994; Pasch et al., 2003; Fischer, 2006) or (Urgelles-Coll, 2010).

We follow the convention established by the PDTB (Penn Discourse Tree Bank, an English corpus annotated at the discourse level (PDTB Group, 2008)), which uses the term *Arg2* for the argument that is linked to the syntactic host clause of the connective, and *Arg1* for the other argument. To make the arguments easy to identify, the span of text corresponding to Arg2 is henceforth in italics, while Arg1 is in boldface; and we use colors for the discourse connectives — see (4).

(4) **Fred didn't go to work** because *he is sick*.

According to (Danlos et al., 2018), the main morpho-syntactic categories of primary connectives are: subordinating or coordinating conjunctions, prepositions or postpositions (in some languages such as German), suffixes (in some languages such as Turkish), adverbs and frozen prepositional phrases. On the semantic side, the PDTB distinguishes around thirty different sense tags of connectives (organized in a hierarchy), which characterize the discourse relation between the arguments of connectives. For example, the sense tag of *because* in (4) is Reason, which applies when the connective indicates that the situation specified in Arg2 is interpreted as the cause of the situation specified in Arg1.

Lexicons for (primary) connectives have first been built for French — LexConn (Roze et al., 2012; Danlos et al., 2015) — and German — DiMLex (Stede, 2002; Scheffler and Stede, 2016). As the DiMLex XML format has proven to be quite compatible with approaches to lexicons in other languages, the original format of LexConn was converted to the DiMLex format, and new lexicons for Italian (Feltracco et al., 2016), Portugese (Mendes et al., 2018) and English (Das et al., 2018) have recently been constructed following the DiMLex format.

3 Templates for secondary connectives

Secondary connectives play the same discourse function as primary connectives, but they are compositional word combinations which are lexically headed by a unique core unit. They often allow modifications and variants as it is the case of verbal MWEs for example.

We describe below four templates for secondary connectives, specifying for each one the syntactic category of its lexical head (which is underlined). These templates are illustrated with English examples, but there exist equivalents in many other languages. As the notion of a secondary connective is relatively new, we are not in a position to guarantee exhaustivity neither to give any frequency data.

3.1 Adverbial prepositional phrases

Secondary connectives in the form of prepositional phrases (PPs) are of two types according to the lexical head: the preposition or the noun. The first type is a combination of a preposition, the lexical head, and an anaphoric expression, mostly a demonstrative pronoun, like *despite/besides this*, *due/thanks to this*, *because/in spite of this*, as in (5a); no modification is possible. The second type is lexically headed by the noun, like *for this reason*, *under these conditions*, *for this purpose*, as in (5b); the noun can be modified by an adjective (*for this unbelievable reason*) and inflected (*for these reasons*).

(5) a. **I had all the necessary qualifications.** Despite this, *I didn't get the job.*

b. **We were stuck in a traffic jam.** For this reason, *we couldn't attend the event.*

These two types of secondary PPs are schematized as PP/Prep and PP/N, respectively: in these schemes, the core unit (Prep or N) is indicated on top of the syntactic category PP.

3.2 Templates headed by a discourse verb

Other secondary connectives are lexically headed by a discourse verb (see note 2) , e.g., *cause*, *precede*, *follow*, *prove* or *mean*. The subject of a discourse verb in a secondary connective is an anaphoric pronoun referring to Arg1. The examples in (6) show that Arg2 can be nominal or clausal.

(6) a. **The increase in MHPG in hypothalamus and brain stem occurred as early as 1 hr postdosing;** this preceded *the earliest measurable sign of tremor and initial hypothermia.* [4]

b. **The student feels more supported and is less afraid to ask.** This provokes that *he does not disconnect and continues to pay attention in class.*[5]

Discourse verbs can be used in the active or passive form (e.g., *this was preceded by, this was caused by*). In the passive form, they signal the "dual relation" of that signalled in the active form, e.g., *precede* in the active form signals a temporal relation in which the situation in Arg1 precedes that in Arg2, and the dual temporal relation in the passive form, i.e. Arg2 precedes Arg1. Discourse verbs can be modified by an adverb with a modal or evaluative value (*this was obviously due to, this unfortunately caused*). They are schematized as DVs.

3.3 Copula templates

Other secondary connectives contain a semantically weak verb, mostly the copula *be*. The lexical head may appear before or after the copula. First, the copula can be built with a subject whose head noun is the lexical head of the connective like *reason*, *condition*, *consequence*, *example*, *conclusion*. The examples in (7) show that Arg2 can be nominal or clausal. The copula can be omitted to yield examples such as (8).

(7) a. **The tourism industry has grown over the years.** The reason is *the arrival of international flights to the capital.*

b. **The tourism industry has grown over the years.** The reason is that *international flights started arriving at the capital.*

(8) **International flights arrived at the capital.** Result: *the tourism industry has grown over the years.*

Second, the copula can be built with a subject which is an anaphoric pronoun referring to Arg1. The lexical head follows the copula: it can be a noun as in (9a), a subordinating conjunction as in (9b), or a preposition as in (9c).

[4] P.H Chen, *Toxicology and applied Pharmacology*, vol 77.

[5] Florentino Blazquez Entonado & Santiago Marin Garcia, *Co-operative learning in the teaching of mathematics in secondary education*.

(9) a. **International flights arrived at the capital.** That is the reason why the tourism industry has grown over the years.

b. **Out in space, the sky looks black, instead of blue.** This is because *there is no atmosphere.*

c. **Jane got pregnant.** This was before *her father's death.*

These copula templates are schematized as BE/SubjN when the lexical head is the head of the subject and as BE/AttN, BE/Conj or BE/Prep otherwise. When the core unit is a noun, it can be modified by an adjective (*a possible reason is that, that is the simple reason why*). The copula can be modified by an adverb (*this was probably before*).

It should be noted that the noun *reason* is the lexical head of the secondary connective both in (7) and (9a), but it expresses that Arg2 is the reason of Arg1 in the former examples and the dual relation in the latter.

3.4 Secondary subordinating conjunctions and prepositions

This set of secondary connectives includes PPs that may introduce a complement in the form of a clause, a VP or an NP, see (10). They are in the form *Prep the N that/of* where *N* is the lexical head, which can be modified (*in the vain hope that*). They can be qualified as secondary subordinating conjunctions or prepositions.[6]

(10) a. **People were trained** in the hope that *they would find jobs.*

b. **Rwanda is also developing ICT** with the aim of *becoming a regional hub and supporting economic growth.*

c. **He applied for a job in a new city** in the hope of *a positive answer.*

These PPs may appear without a complement but with an anaphoric determiner, as in *Many of its operational programmes and activities have been reoriented in line with this aim.*

4 Use of secondary connectives in discourse annotation and parsing

In discourse annotation, the first task is to identify the primary connectives in the corpus. This task generally relies on a lexicon of primary connectives which is projected into the corpus. A verification phase is needed to check that any item in the lexicon is actually used as a connective in the corpus: for example, *in short* in English is either a connective adverbial or a word sequence in the MWE *in short supply*; similarly *bref* in French is either a connective adberbial ('in short') or an adjective ('short'). Next, the sense and the arguments of the connective are annotated.

During the annotation of the PDTB (Prasad et al., 2008), it has been noted that discourse relations may be explicitly signalled by word combinations which are not primary connectives; they are called "alternative lexicalizations" (AltLex) (Prasad et al., 2010). In the PDTB-2 corpus, 624 tokens are annotated as AltLex. They correspond in our terminology either to primary connectives[7], or secondary connectives or free connecting phrases. According to (Prasad et al., 2010), word combinations are annotated as AltLex when there is no (primary) connective and when "a discourse relation is inferred, but insertion of a connective leads to redundancy". We believe that redundancy is not a clear criterion for identifying AltLex, because a discourse segment can be introduced by both a primary and a secondary connective that signal the very same discourse relation, as in the real-life examples (11) below. In (11a), the result discourse relation is signalled by both *as a result* and *this caused*; when following the PDTB annotation guidelines, *this caused* would not be annotated as an AltLex, because of the presence of the primary connective *as a result*. This leads to incoherencies and to wrong identification of arguments: the Arg2 of

[6] As far as we know, there are no multi-word expressions that can be qualified as secondary coordinating conjunctions.

[7] A primary connective may be annotated as an AltLex because the list of primary connectives in the PDTB-2 is made up of only 100 elements, and so several connectives are missing, like the adverb *thereafter,* as well as any preposition with a discourse use, like *in order to*. The English lexicon of (Das et al., 2018) contains 150 entries.

as a result would be wrongly identified as the rest of the sentence, i.e. *this caused families to send their children to work*, while it is rightly identified as *families to send their children to work* when *this caused* is recognized as a secondary connective.

(11) a. **Families, especially in Lebanon, have passed through different decades of wars (...).** As a result, this caused *families to send their children to work.*[8]

b. **... the patient began to show evidence of insanity by incoherent talk, false ideas, nervousness, and outbursts of vicious excitement.** Later, this was followed by *mutism, refusal to eat, and stupor.*[9]

Moreover, a discourse segment can be introduced by both a primary and a secondary connective that signal different discourse relations, as in (12) in which *however* signals a contrastive relation and *this caused* a causal relation. Repairing these two connectives leads to the right identification of Arg2 (in italics).

(12) **The Middle Unit of the Chalk Mar was encountered in the crown of the TBM drives on the marine side of the NATM excavations.** However, this caused *only very limited delays.*[10]

In conclusion, we believe that the first task in discourse annotation should be to identify both primary and secondary connectives in the corpus, thanks to the projection of lexicons for the two types of connectives. As far as we know, no verification phase is needed after projecting the lexicon of secondary connectives into a corpus, in contrast with primary connectives (see above).[11] We underline that both primary and secondary connectives can be embedded under a report or propositional attitude verb, as in (13), which means that there is no *a priori* constraint, when projecting lexicons into a corpus, on the position of connectives in a sentence.

(13) a. **Fred will go to Peru next year.** Jane thinks on the other hand *he will go to France.*

b. Because of this, **the bitcoin address could be well formed . . . in a sense . . .** I suppose this is the reason why *blockchain doesn't reject it.*[12].

So far, we have only discussed cases where discourse relations between two spans of text are explicitly signalled by lexical items (primary or secondary connectives, or free connectives phrases); such relations are called "explicit relations". However, it happens quite frequently — roughly, half of the time (Braud and Denis, 2016) — that discourse relations are not overtly marked; they are called "implicit relations". To illustrate, the causal relation between Fred's jokes and his friends' hilarity is explicit in (14a-b) — first presented in Section 1 — and implicit in (14c)

(14) a. **Fred didn't stop joking.** As a result, *his friends enjoyed hilarity throughout the evening.*

b. **Fred didn't stop joking.** This caused *hilarity among his friends for the whole evening.*

c. **Fred didn't stop joking.** *His friends enjoyed hilarity throughout the evening.*

Implicit relations are hard to handle both in discourse annotation and parsing. In discourse annotation, the annotators are supposed to infer the type (sense) of the implicit relations, which leads to a decrease of the inter-annotator agreement in comparison with the annotation of explicit relations. In discourse parsing, the situation is even worse: in shallow discourse parsing, which aims at automatically identifying discourse relations and their arguments in text, determining the type of relation is much more difficult

[8]Philippe W. Zgheib, *Business & Economics*.

[9]Bernard Glueck, *Studies in Forensic Psychiatry*.

[10]Colin S. Harris, Paul M. Varley, Colin D. Warren, *Technology & Engineering*.

[11]This also contrasts with MWEs : a MWE recorded in a lexicon may have a literal meaning in a corpus and not its idiomatic one (Savary and Cordeiro, 2018), while a secondary connective has only a literal compositional meaning.

[12]`https://bitcoin.stackexchange.com/questions/67126/is-1wh4bh-a-valid-bitcoin-address`

for implicit than for explicit relations, e.g., in the system of (Oepen et al., 2016), the difference in F1-measure is 13 points. Essentially all shallow discourse parsers (see the shared tasks at CoNLL 2015 and 2016) follow the pipeline model implemented in (Lin et al., 2014), which first identifies connectives, their arguments, and the relations they signal, and in a later stage tries to classify implicit relations using a separate module.

As both discourse annotators and parsers first identify connectives, and next try with difficulty to handle implicit relations, it is better to reduce the number of implicit relations as much as possible. This can be done if the expressions we have labeled secondary connectives are recognized as connectives. For example, if the expression *this caused* in (14b) is identified, thanks to a lexicon, as a secondary discourse connective with a causal meaning, no implicit relation between the two sentences is at stake and the causal relation can be identified as easily as in (14a).[13]

In conclusion, we believe that systematically handling secondary connectives — and having them play a similar role to that of primary connectives — can be highly beneficial both for discourse annotation and parsing.

5 Building lexicons for secondary connectives

We present two solutions to organize a lexicon for secondary connectives: the first one, which relies on the templates in which secondary connectives appear, facilitates the projection of secondary connectives into corpora — which helps both discourse annotation and parsing, as shown in the previous section — but may be hard to use for a human reader; the second one, which relies on the lexical heads of secondary connectives, is hard to use for projection into corpora but easy to read for a human user. The first solution is explored for a French lexicon, the second one for a Czech lexicon (Mírovský et al., 2017), which is under development, following annotation of the PDiT (Prague Discourse Treebank 2.0 (Rysová et al., 2016)).

The template-based solution consists in dividing the secondary connective lexicon into sub-lexicons, a sub-lexicon per template. This means creating a sub-lexicon for the template PP/Prep, another one for the template PP/N, another one for the template DV, etc. This solution can be compared to what is done for MWEs: there exist sub-lexicons for verbal, nominal, adjectival, ... MWEs. In a sub-lexicon for a given template, the lexical entries are the lexical heads which are all of the same syntactic category, e.g. *Prep* in the sub-lexicon of the secondary connectives which follow the PP/Prep template or *V* in the sub-lexicon for the DV template. The properties recorded in a sub-lexicon may describe the possible variants and modifications of secondary connectives — e.g. in the DV sub-lexicon, the possibility for the verb to be used in the passive form, or in the PP/N sub-lexicon, the possibility for the noun to be modified and/or inflected. Other properties may describe the sense(s) of the secondary connective(s)[14], the existence of an equivalent primary connective if any, etc.

This organization in template-based sub-lexicons facilitates the projection of secondary connectives into a corpus since each template corresponds to one or two regular expression(s) which define search patterns to locate the secondary connectives into a corpus. However, this solution can be inconvenient for a human user, who may want to find all the secondary connectives with the same lexical head in one place, as advocated in (Danlos et al., 2018). So another solution to build a lexicon of secondary connectives is to keep all the secondary connectives with the same lexical head within a single entry. Such a lexicon is illustrated for English in (15) for the entry *reason*, which is the lexical head of secondary connectives following three templates. For each template, its scheme (e.g., PP/N) is given and followed by its specification. The following abbreviations are used in the specifications of the schemes: Ana for anaphoric, Det for determiner, Adj for adjective, Pro-Subj for a subject pronoun (referring to an eventuality); the symbol N stands for a variable whose value is given in the lexical head field; optionality is marked with parenthesis and alternatives are written within brackets. Under each scheme, the field

[13] In the discourse annotation/parsing of (14a-b), the question left after identification of the primary or secondary connective is to find the first argument (Arg1) of the causal relation. It boils down to find an antecedent to the anaphora in (14b).

[14] In the DV sub-lexicon, two senses (corresponding to a discourse relation and its dual relation) must be recorded if the discourse verb can be used in the passive form (Section 3.2).

Realizations gives concrete examples of the scheme, and the binary feature Inflection indicates if N can be inflected.

(15) **Lexical head: N = *reason***

Scheme 1 = PP/N : for [Ana-Det (Adj)/Ana-Adj] N
Realizations: *for this reason*, *for given reason*
Inflection = 1

Scheme 2 = BE/SubjN: Det (Adj) N BE (that)
Realizations: *the reason is*, *a possible reason is that*
Inflection = 1

Scheme 3 = BE/AttN: Pro-Subj BE the (Adj) N why/for which
Realizations: *that is the reason why*; *this is the simple reason for which*
Inflection = 1

The information given in (15) is complemented by other fields that appear in each scheme: for example, the sense of the secondary connective, the existence of a primary connective equivalent if any, foreign language equivalents, etc. This is illustrated in (16) for the first scheme.

(16) **Lexical head: N = *reason***

Scheme 1 = PP/N : for [Ana-Det (Adj)/Ana-Adj] N
Realizations: *for this reason*, *for given reason*
Inflection: 1
Sense : Result
Primary connective equivalent: therefore
Foreign language equivalents:

- Czech: z tohoto důvodu
- French: pour cette raison
- German: aus diesem Grund

Scheme 2 = BE/SubjN: Det (Adj) N BE (that)
...

6 What's else could lexically signal discourse relations?

So far, we have seen that discourse relations can be lexically signalled by primary connectives, secondary connectives or free connecting phrases. Are they the only word combinations used to signal a discourse relation between two spans of text? The answer is no: there are other devices, for example constructions such as the comparative correlative construction (CC) discussed *inter alia* by (Ross, 1967; McCawley, 1988; Fillmore, 2013) and illustrated in (17a), or the reversed CC illustrated in (17b) and whose interpretation is close to that of the sentence in (17c) with a primary connective.

(17) a. The more I read, the more I understand.

b. I understand more, the more I read.

c. **I understand more** as *I read more,*

We call "discourse constructions" examples such as (reversed) CCs in (17a-b), with the aim of filling the gap between the domains of discourse studies and construction grammars. These two domains ignore each other most of the times, though not always. For example, conditional sentences with the primary connective *if* have long been studied in construction grammars (Jackendoff, 2002). On the other side, discourse studies are concerned with "parallel connectives" which are defined in (PDTB Group, 2008) as "pairs of connectives where one part presupposes the presence of the other, and where both together take the same two arguments", and are illustrated in (18); parallel connectives share the same structure as that of CCs, namely: *X1 S1 Punct X2 S2*.

(18) a. On the one hand, Mr. Front says, it would be misguided to sell into "a classic panic." On the other hand, it's not necessarily a good time to jump in and buy.

b. If the answers to these questions are affirmative, then institutional investors are likely to be favorably disposed toward a specific poison pill.

We wish to add that some some lexemes that usually are primary connectives, such as the subordinating conjunction *because*, should not be treated as such in (19), as discussed in construction grammars (Fillmore, 2013).

(19) Just because I live in Berkeley doesn't mean I'm a revolutionary.

In a nutshell, there exist grammatical constructions which should be discussed in discourse studies. This is left for future research.

7 Conclusion

We have shown that discourse relations can be lexically signalled by primary connectives, secondary connectives, free connecting phrases, and also by some discourse constructions. Primary connectives are often grammaticalized variants of secondary connectives, and there already exist primary connective lexicons in various languages, which describe the properties from 150 to 250 lexical entries (lexemes or MWEs). Secondary connectives, which are compositional, modifiable and inflectable, appear in syntactic structures that are lexically headed by a unique core unit, i.e., the unit that has the strongest meaning. It is possible to build lexicons for secondary connectives, and we have presented two ways to organize such lexicons, one primarily based on syntax, the other on lexical heads. Free connecting phrases are compositional and include at least two content words, which make it undesirable to record them in a lexicon. Discourse constructions need further work.

We explained why primary and secondary connective lexicons are quite helpful for both discourse parsing and annotation. The reasons are twofold: right identification of the Arg2 of discourse relations and decrease of number of implicit relations. These reasons are practical and not theoretical. Nevertheless, we want to emphasize the following point : in discourse theories (such as RST or SDRT (Mann and Thompson, 1988; Asher and Lascarides, 2003)), which rely on compositional semantic analyses without making use of the notion of secondary connectives, an implicit relation would be posited in (14b) for example, but what is the type of this discourse relation? It can be only a very weak relation such as "Commentary", which means that the second sentence gives a comment on/ a follow up of the first one. Such a relation doesn't give a structural analysis of discourse, which is the aim of RST and SDRT, and thus should be avoided anyway. Moreover, we have no idea on how these theories would handle discourse constructions such as those briefly discussed in Section 6.

One question is left open: we have shown that lexicons for secondary connectives can be built although these word combinations are compositional — but lexically headed. Is it envisageable to build other lexicons recording compositional lexical entries? And if it is, for which type of lexical entries?

Acknowledgements

I thank Marie Candito and the reviewers for their fruitful comments.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Nicholas Asher. 1992. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 203–213, Austin, Texas.

Laurence Danlos, Margot Colinet, and Jacques Steinlin. 2015. FDTB1 : Repérage des connecteurs de discours dans un corpus français. *Revue Discours*, 15.

Laurence Danlos, Katerina Rysov, Magdalena Rysov, and Manfred Stede. 2018. Primary and secondary discourse connectives: definitions and lexicons. *Dialogue & Discourse*, 9-1:50–78.

Laurence Danlos. 2006. Discourse verbs and discourse periphrastic links. In *Proceedings of the second workshop on Constraints in Discourse (CID 2006)*, Maynooth, Ireland.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2018)*, Melbourne, Australia.

Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. Lico: A lexicon of Italian connectives. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-IT 2016)*, Napoli, Italy.

Charles Fillmore. 2013. Berkeley Construction Grammar. In Thomas Hoffmann and Graeme Trousdale, editors, *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

Kerstin Fischer. 2006. Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. *Approaches to discourse particles*, pages 1–20.

Josef Hrbáček. 1994. *Nárys textové syntaxe spisovné češtiny*. Trizonia, Prague, Czechia.

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Neville Ryant Karin Kipper, Anna Korhonen and Martha Palmer. 2006. Extensive classifications of English verbs. In *Proceedings of the 12th EURALEX International Congress*, Turin, Italy.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

Gyri Smrdal Losnegaard, Federico Sangati, Carla Parra Escartn, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on mwe resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

James McCawley. 1988. The comparative conditional construction in English, German and Chinese. In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 176–187.

Amalia Mendes, Iria del Rio Gayo, Manfred Stede, and Felix Dombek. 2018. A lexicon of discourse markers for Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Jiří Mírovský, Pavlína Synková, Magdaléna Rysová, and Lucie Poláková. 2017. *CzeDLex 0.5*. Charles University, Prague, Czech Republic.

S. Oepen, J. Read, T. Scheffler, U. Sidarenka, M. Stede, E. Velldal, and L. vrelid. 2016. OPT: OsloPotsdamTeesside—Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing. In *Proceedings of the CONLL 2016 Shared Task*, Berlin, Germany.

Renate Pasch, Ursula Brauße, Eva Breindl, and Ulrich Herrmann Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

PDTB Group. 2008. The Penn Discourse Treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science, University of Philadelphia.

Rashmi Prasad, Nikhil Dinesh, Alan Leea, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010): Poster Volume*, Beijing, China.

John Robert Ross. 1967. *Constraints on Variables in Syntax*. PhD. MIT, USA.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LexConn: a French Lexicon of Discourse connectives. *Revue Discours*, 10.

Magdaléna Rysová and Kateřina Rysová. 2014. The centre and periphery of discourse connectives. In Wirote Aroonmanakun, Prachya Boonkwan, and Thepchai Supnithi, editors, *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC 2014)*, pages 452–459, Bangkok, Thailand.

Magdaléna Rysová, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, and Šárka Zikánová. 2016. *Prague Discourse Treebank 2.0*. Charles University, Prague, Czech Republic.

Agata Savary and S R. Cordeiro. 2018. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT 16)*, Charles University in Prague, Czech Republic.

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Manfred Stede. 2002. DiMLex: A lexical approach to discourse markers. In A. Lenci and V. Di Tomaso, editors, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Miriam Urgelles-Coll. 2010. *The syntax and semantics of discourse markers*. Continuum Studies in Theoretical Linguistics. A&C Black, London, UK.

Arnold M. Zwicky. 1985. Clitics and particles. *Language*, 61(2):283–305.