Dialogue & Discourse 9(1) (2018) 50–78 doi: 10.5087/dad.2018.102

# Primary and secondary discourse connectives: definitions and lexicons

**Laurence Danlos** LAURENCE.DANLOS@LINGUIST.UNIV-PARIS-DIDEROT.FR
*Université Paris Diderot, Laboratoire de Linguistique Formelle*
*2 Place Thomas Mann, 75013 Paris, France*

**Kateřina Rysová** RYSOVA@UFAL.MFF.CUNI.CZ
*Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics*
*Malostranské náměstí 25, 118 00 Praha 1, Czech Republic*

**Magdaléna Rysová** MAGDALENA.RYSOVA@UFAL.MFF.CUNI.CZ
*Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics*
*Malostranské náměstí 25, 118 00 Praha 1, Czech Republic*

**Manfred Stede** MANFRED.STEDE@UNI-POTSDAM.DE
*Universität Potsdam, Applied Computational Linguistics Discourse Research Lab*
*Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany*

**Editor:** Demberg Vera

Submitted 06/2017; Accepted 04/2018; Published online 06/2018

## Abstract

Starting from the perspective that discourse structure arises from the presence of coherence relations, we provide a map of linguistic discourse structuring devices (DRDs), and then focus on those found in written text: connectives. To subdivide this class further, we follow the recent idea of structuring the set of connectives by differentiating between primary and secondary connectives, on the one hand, and free connecting phrases, on the other. Considering examples from Czech, English, French and German, we develop definitions of these groups, with attention to certain cross-linguistic differences. For primary and secondary connectives, we propose that their behavior can be described to a large extent by declarative lexicons, and we demonstrate a concrete proposal which has been applied to five languages, with others currently being added in ongoing work. The lexical representations can be useful both for humans (theoretical investigations, transfer to other languages) and for machines (automatic discourse parsing and generation).

**Keywords:** discourse connective, discourse structure, discourse semantics

## 1. Introduction

An important strand of research on discourse coherence operates under the assumption that *discourse relations* are, in addition to coreference, of central importance to explain local coherence between sentences, or more generally, between discourse segments of various kinds. Discourse theories and annotation projects implementing this insight have been proposed, *inter alia* by Mann and Thompson (1988); Sanders et al. (1992); Asher and Lascarides (2003); Prasad et al. (2008). While the precise sets of relations proposed (and the underlying motivations) vary, there is general agreement on coarse classifications into groups such as *causal*, *contrastive*, and *additive*. Following the notion of discourse relations further, it is evident that their instances in actual text or speech can

be either *signalled* or *unsignalled* (implicit). In the former case, some linguistic device indicates to the hearer or reader the presence of the discourse relation — sometimes rather specifically (e.g., *although*), sometimes more vague (e.g., *and*).

These devices are available in all the world's languages, but they can be of quite different kinds, and so we start here by using the very general term "discourse relational devices" (DRDs) for them. This follows the practice of the European *TextLink* COST action[1], which aims at inventarization, annotation, and cross-linguistic comparison of DRDs. The research reported in this paper is a result of a collaboration within this project.

The set of DRDs can first be divided in two subsets: discourse *connectives*, on the one hand, and discourse *markers,* or *particles,* on the other. Discourse connectives establish a *two-place* relation between the 'abstract objects' (Asher, 1993) that result from interpreting the text spans related by the connective, called *arguments* of connectives. On the other hand, discourse markers/particles establish a *one-place* relation; as stated in (Fischer, 2000, p. 1), they are "small items such as German *ja, also, ne, oh* or *ach* and English *yes, yeah, oh* or *well* which predominantly occur in spontaneous spoken language." We acknowledge that the distinction is somewhat blurred by the fact that connectives trivially are also used in spoken language, and - more interestingly - some connectives in some languages have additional senses such that they can also function similar to discourse markers. Still, we see the difference between two-place and one-place relations as an important categorizing factor.

In this paper, we look only at discourse connectives, in written texts. We seek to provide further meaningful subclassifications for them. Connectives do not form a syntactically homogeneous category, and this is one reason why existing grammars and dictionaries fail to adequately capture the meaning, usage conditions, and translation options of discourse connectives. But this knowledge is of great importance, *inter alia* for supporting foreign language learning, and for building computer software that performs applications involving text understanding. Consequently, we see a pressing need for defining inventories of discourse connectives in the form of lexicons that are readable both for humans and for machines. Our goal here is to propose guidelines for developing such lexicons, which can be applied to different languages, resulting in resources that can be interlinked across languages.

As part of a definition of discourse connectives, we propose to make a distinction between *primary* and *secondary* discourse connectives, illustrated below in (1a) and (1b), respectively.[2] In (1a), the causal relation between Fred's jokes and his friends' hilarity is explicitly signalled by the primary connective *as a result,* which is a frozen multi-word unit, while, in (1b), it is signalled by the expression *this caused*, which is made of a subject and a verb, both used with a compositional meaning. The paraphrase relationship between the examples demonstrates that *this caused*, which we will call a secondary connective, can be used to signal a causal relation between two events in the same conditions as the phrase *as a result* (as long as syntactic constraints are respected).

(1) 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.

(2) Fred didn't stop joking. His friends enjoyed hilarity throughout the evening.

1. http://textlink.ii.metu.edu.tr
2. In this paper, most of the time, we deliberately use invented examples for purposes of succinctly illustrating particular phenomena or distinctions.

Another distinction, already mentioned above, is usually made between *explicit* and *implicit* discourse relations, depending on whether they are lexicalized by a connective or not. To illustrate, the causal relation between Fred's jokes and his friends' hilarity is not overtly marked in (2) and is said to be implicit; it is said to be explicit in (1a), thanks to the primary connective *as a result*. This difference is a particularly acute phenomenon in *shallow discourse parsing*, which aims at automatically identifying discourse relations and their arguments in text. For current parsers, determining the type of relation is much more difficult for implicit than for explicit relations (e.g., in the system of Oepen et al. (2016), the difference in F1-measure is 13 points; see also (Braud and Denis, 2016)).

As far as examples like (1b) are concerned, standard discourse parsers would not recognize the expression we have labeled as a secondary connective here, and would thus posit an implicit relation between the two sentences.[3] Depending on the models used, *caused* may happen to have been learned as a cue for the statistical classifier, but in general the causal relation between Fred's jokes and his friends' hilarity will be hard to detect. On the other hand, if the expression *this caused* were identified by a declarative resource as a secondary discourse connective with a causal meaning, the relation could be identified as easily as in (1a). Therefore, we believe that a systematic treatment of secondary connectives — and having them play a similar role to that of primary connectives — can be highly beneficial for such parsers, because delegating cases like (1b) to the "explicit" module would reduce the number of instances to be handled by the much more ambitious "implicit relation" module.

In a nutshell, the two central contributions of this paper are

- to revisit and reconsider the earlier work that previous authors have done on lexicons of discourse connectives, adding insights gained from a multilingual perspective that was originally not present, and thus providing a more comprehensive landscape of DRDs/connectives;
- to formulate guidelines for building such lexicons in other languages, drawing from our experiences in both monolingual description and interlingual linking.

The paper is organized as follows. Section 2 provides an overview of the types of discourse connectives and some related DRDs. We characterize the difference between primary and secondary connectives, which was originally introduced by Rysová and Rysová (2014). These two categories are then defined more precisely in Sections 3 and 4, which in turn suggest further subgroups and discuss various details. Our descriptions are based on Czech, English, French and German, but we give examples in English whenever possible illustrating phenomena that are common to these four languages (and possibly other languages). Furthermore, both of these sections present our suggestions for developing new connective lexicons: how to find the set of connectives, what information to encode about them, and in which format. For primary connectives, they build on our experiences with the two lexicons 'DiMLex,' for German, (Stede, 2002; Scheffler and Stede, 2016) and 'LexConn', for French (Roze et al., 2012; Danlos et al., 2015). For secondary connectives, our exposition is based on the work described in Rysová (2015) for Czech, which is applied in the 'CzeDLex' lexicon (Mírovský et al., 2017). In Section 5, we add observations on dealing with the semantics of the two categories of connectives and point to the need for further work in this regard. Finally, Section 6 draws general conclusions.

3. 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 relation, and in a later stage tries to classify implicit relations using a separate module.

## 2. Connectives and related DRDs

This section describes our terminology and notation, surveys the realm of connectives and other, similar DRDs, and argues that two groups (primary and secondary connectives) should be subject to a lexical description.

### 2.1 Connectives: primary, secondary, and free phrases

Due to its syntactic heterogeneity, the notion of discourse connective needs to be defined at the functional level: a discourse connective is an expression whose function is to semantically and/or rhetorically link two "abstract objects" in the terminology of Asher (1993) — roughly, events, states or propositions. This functional definition means that a discourse connective is a discourse-level predicate with two arguments. 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 an Arg2 is henceforth set in italics, while Arg1 is in boldface; and we use colors for the discourse connectives — see (3).

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

b. **Fred is sick**. However, *he did go to work*.

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 (3a) 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. Senses of connectives will be discussed again in Section 5.

Turning again to syntax, there are several types of DRDs that explicitly express discourse relations. To illustrate, in (4a-e), there is a common relation of Result expressed by different linguistic means, ranging from a grammaticalized one-word expression (*therefore*) to an open combination of words (*due to this weather*). These expressions differ in many ways, and we divide them into three groups following the concept of Rysová and Rysová (2014): primary connectives (4a), henceforth in magenta; secondary connectives (4b-d), in blue; and free connecting phrases (4e), underlined.

(4) **There was nice spring weather with high temperatures already in April.**

a. Therefore, *domestic productions were arriving in the market much earlier than normally.*

b. Because of this, *domestic productions were arriving in the market much earlier than normally.*

c. For this reason, *domestic productions were arriving in the market much earlier than normally.*

d. This caused *domestic productions to arrive in the market much earlier than normally.*

e. Due to this weather, *domestic productions were arriving in the market much earlier than normally.*

Primary connectives are single-word units or non-compositional multi-word units. Secondary connectives are multi-word units that differ from primary connectives in that they are not (yet) fully grammaticalized. For example, a secondary connective can be composed of a preposition, e.g., *because of,* followed by a demonstrative pronoun with a compositional anaphoric reading, e.g., *this*, (4b). Another example of secondary connectives is illustrated in (4c) with the expression *for this reason,* which is compositional and allows internal modification and inflection *(for this unbelievable reason, for these reasons, for that reason)*. In (4d), it is the verb *cause,* with an anaphoric subject referring to Arg1, which expresses the causal relation.

Secondary connectives are distinct from free connecting phrases, such as *due to this weather* in (4e), in that the former can be used in a wide variety of contexts while the latter can be used only in quite specific contexts. As an illustration, the expression *because of this illness* in (5a) is a free connecting phrase, whose 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 (5b) — which is incoherent (hence the sign #) — because there is no antecedent for the anaphoric noun phrase. It contrasts with the expression *because of this*, which can be used in either context, (5c).

(5) 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.*

### 2.2 The notion of 'alternative lexicalization'

The Penn Discourse Treebank (Prasad et al., 2008) employs the notion of *alternative lexicalization* (AltLex) for expressions that are not considered connectives, yet explicitly signal the presence of a relation. In our terminology, these correspond to the two groups of secondary connectives and free connecting phrases. In the PDTB-2 corpus, 624 tokens are annotated as AltLex.[4] According to Prasad et al. (2010), expressions are annotated as AltLex when "a discourse relation is inferred, but insertion of an implicit connective leads to redundancy" (Prasad et al., 2010). We feel that redundancy is not a very clear criterion for identifying AltLex, because an Arg2 can easily be introduced by both a primary and a secondary connective that signal the very same discourse relation, as in the real-life example (6) below, in which Result is signalled by both *as a result* and *this caused*.[5] When following the PDTB annotation guidelines, *this caused* would be annotated as an AltLex for instance in (1b) but not in (6), because the explicit connective *as a result* is present. Examples such as (6) show that (i) redundancy is observed in real texts, and (ii) the definition/annotation of AltLex should not rely on redundancy but on semantics: an AltLex (a secondary connective or free connecting phrase in our terminology) is an expression that has the same function as a primary connective but has different distributional properties.

4. Some primary connectives are also annotated as AltLex. This happens 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*.

5. Source: Business Ethics and Diversity in the Modern Workplace Philippe W., Zgheib publisher, 2014

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

### 2.3 Relative pronouns

Primary/secondary connectives and free connecting phrases are not the only types of DRDs that can connect two spans of text; relative pronouns can also serve this purpose. Whereas *restrictive* relative clauses merely work as modifiers of NPs (serving to identify the referent from a set of candidates), *non-restrictive* relative clauses can have discourse-structural roles. For example, the relative pronoun in (7a) does not modify its antecedent but is used to connect the eventualities of Ted's giving and Mary's eating. See also (7b) from (Huddleston and Pullum, 2002, page 1223) who speak of the "continuative use of supplementary relatives". Along the same lines, *which* in (7c) can be considered as connecting two eventualities.

(7) a. Ted gave a candy to Mary, who ate it quickly.

b. I gave it to John, who passed it on to Mary, and she gave it back to me.

c. Ted gave a candy to Mary, which pleased her.

It is not appropriate to include relative pronouns in discourse connective lexicons: they are a separate class. This, of course, does not preclude useful work on relative clauses. For example, further work is needed in order to clarify the linguistic contexts where relative pronouns function like DRDs, and how annotate them as DRDs in corpora. Indeed, some sentences including a relative pronoun in one language may translate as two sentences linked by a connective in another language. For instance, see (8) from the English-French Hansard corpus.

(8) a. You could call it "Canada's moment", a chapter in history that finds the Canadian economy outperforming expectations while major foreign economies are still in distress.

b. **On peut dire que le Canada est en train de vivre son heure de gloire.** En effet, *l'économie canadienne dépasse les attentes alors que les principales économies étrangères sont encore en pleine crise.*

### 2.4 Lexical description

In conclusion, among the set of connective-like DRDs, we first have identified the groups of primary and secondary discourse connectives. These will be discussed in detail in the next two sections. We argue that each group, despite the *prima facie* heterogeneity of the items included, can be described by a lexical resource that gathers the information on the commonalities and differences between these items. Such lexicons can support empirical analyses and theoretical work, and they can assist automatic discourse parsers in identifying connectives and their arguments in text. For primary connectives, a lexical approach is feasible because they are closed-class items with, as we will see, a relatively small membership. Secondary connectives will be described by templates that are, in a sense, open, yet they are headed by a set of items that again we take to be closed.

Free connecting phrases, on the other hand, are not amenable to a lexical description because of their productive nature. "Lexical entries" like *because of this (bad) illness* or *due to this weather* would be undesirable because of their heavy dependence on context. For these reasons, we do

not believe that free connecting phrases are *lexical* items like primary and secondary connectives are, and they will therefore be excluded from further consideration in this paper. Likewise, we have briefly mentioned relative pronouns above, which share some properties with connectives, but should be treated separately.

## 3. Primary discourse connectives

The previous section leads us to define a primary discourse connective as an element which is, morpho-syntactically, a frozen single-word unit or a non-compositional multiword unit, and which is, 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).

This description pertains to a variety of expressions across languages, which are of different syntactic categories, as we have illustrated in the previous section. To organize the following in-depth discussion, the syntactic categories will be described following the distinction between "intra-sentential" and "inter-sentential" connectives in Sections 3.1 and 3.2 below.

In general, intra-sentential connectives typically join both arguments within one typographic sentence (9a), while inter-sentential connectives typically appear in (the second sentence of) a sequence of two typographic sentences, as in (9b) or (9c).

(9) a. **Fred is nice** but *he may be tough with women.*

b. **Fred is nice.** Therefore, *he is never tough with women.*

c. **Fred is nice.** *He is therefore never tough with women.*

However, it is well-known that typographic conventions are only preference rules, which means that (10) can be observed in parallel with (9).

(10) a. **Fred is nice**. But *he may be tough with women.*

b. **Fred is nice**, *he is therefore never tough with women.*

Therefore, we propose a linguistic criterion to help distinguish intra- and inter-sentential connectives:

- **Intra**-sentential connectives form discourse segments that **can** be embedded under a matrix clause, (11a),
- **Inter**-sentential connectives form discourse segments that **cannot** be embedded under a matrix clause, (11b).

(11) a. Jane said that [**Fred is nice** but *he may be tough with women.*].

b. *Jane said that [Fred is nice, he is therefore never tough with women].[6]

6. This example becomes acceptable only if *and* - an inter-sentential connective — is added: *Jane said that [Fred is nice and (he) is therefore never tough with women].*

Intra-sentential connectives (e.g., subordinating and coordinating conjunctions) are thoroughly described and analyzed in syntax/semantics studies that are concerned with the sentential level. Our discourse perspective should be compatible with those studies. The English syntactic terminology we use here largely follows (Huddleston and Pullum, 2002).

Notice that, with respect to these categories, some connectives have more than one use: For example, *before* can be used as an intra-sentential connective (as a subordinating conjunction in (12a) and as a preposition in (12b)), but also as an inter-sentential connective (as an adverb in (12c)).

(12) a. **Fred cooked a pizza** before *Mary finished her homework.*

b. **Fred cooked a pizza** before *finishing his homework.*

c. **Mary finished her homework.** Before, *Fred had cooked a pizza.*

### 3.1 Intra-sentential primary connectives

Intra-sentential connectives are described according to their syntactic category, which puts constraints on the nature of their arguments Arg1 and Arg2, and on their linear order. Recall that Arg2 is, roughly, the content of the host clause of the connective.

#### 3.1.1 SUBORDINATING CONJUNCTIONS

Subordinating conjunctions can be single-word units *(if, because)* or multi-word units (*so that, even if)*. They introduce a finite clause whose mood (indicative or subjunctive in French) or verbal position (V-2 or V-F in German) depends on the conjunction. They are obligatorily positioned in front of the finite clause they introduce, whose content is their Arg2, and they are placed either after their Arg1 (13a), before it (13b), or within it (13c).

(13) a. **John is in a bad mood** because *he lost his keys.*

b. Because *he lost his keys,* **John is in a bad mood.**

c. **John,** because *he lost his keys,* **is in a bad mood.**

Some subordinating conjunctions (but not all) can be externally modified by adverbials such as *only* or *mostly* (focus particles), as in (14). External modification should not be confused with internal modification , which is observed only with multi-word expressions.[7] For example, some multi-word expressions function as subordinating conjunctions in that they introduce a finite clause, e.g., *in the hope that* (15), but are not yet fully grammaticalized, so that internal modifiers can be added *(in the vain hope that)*. They should be considered as "secondary subordinating conjunctions" (Section 4.1.1).

(14) a. Only because *he wanted to see the MOMA,* **Fred went to New York.**

7. In (Prasad et al., 2010), no distinction is made between external and internal modification. Therefore, it is claimed that DRDs "should be treated as an open class that includes explicit connectives". We believe, on the contrary, that primary connectives form a closed class (Section 3.4). As a matter of comparison, English modal verbs (*may, must, can*, ...) can be externally modified (*may possibly, may not*, ...), nevertheless, they do not form an open class.

b. #Only as he wanted to see the MOMA, Fred went to New York.

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

Note that in English *when and if* and *if and when* could be considered multi-word primary conjunctions (PDTB Group, 2008). Along the same lines, in German, it is possible for a conjunction and an adverbial to form a lexicalised (but potentially discontinuous) complex expression that together signal a coherence relation. A case in point is *wenn .. auch*, where *wenn* ('if'/'when') and *auch* ('also') each can signal a different relation by themselves (Condition and Elaboration, respectively), but in combination mark a Concession relation. The words can occur adjacent to one another, usually as an afterthought (16a), or discontinuously (16b).

(16) a. **Dieser Mac ist ziemlich gut,** wenn auch *teuer*.
'This Mac is pretty good, even though expensive.'

b. **Ich nehme diesen Mac,** wenn *er* auch *ziemlich teuer ist.*
'I will take this Mac, even though it's quite expensive.'

### 3.1.2 ADPOSITIONS

Adpositions are prepositions or postpositions which introduce an infinitival or gerund-participial VP (17a and 17b), or an NP referring to an eventuality (17c). In German, adpositions introduce only NPs, except for *um zu (in order to)*.

(17) a. **Fred made a pizza** in order to *please Mary.*

b. **Fred left** after *taking a shower.*

c. **Fred left** after *his shower.*

In English and French, there are only prepositions *(Prep)*, whose positional properties are identical to those of subordinating conjunctions (18a-c). Some prepositions are lexically related to conjunctions: *after* + *clause* and *after* + *Ving*, *in order that* + *clause* and *in order to* + *Vinf*. Some prepositions can be externally modified, such as *mainly in order to* in (18d).

(18) a. **Fred made a pizza** in order to *please Mary.*

b. In order to *please Mary,* **Fred made a pizza.**

c. **Fred,** in order to *please Mary,* **made a pizza.**

d. **Schröder visited China,** mainly in order to *secure trade deals for German industry.*

German also has postpositions, which appear after the element they introduce (Arg2). Additionally, the item *wegen* can be used as either a preposition or a postposition (19a and 19b).

(19) a. Wegen *Marys Ankunft* **gingen wir nicht in ein gutes Restaurant.**
'Because of Mary's arrival, we didn't go to a good restaurant.'

b. *Marys Ankunft* wegen **gingen wir nicht in ein gutes Restaurant.**
'Because of Mary's arrival, we didn't go to a good restaurant.'

"Circumpositions" in German are multi-word connectives made up of a preposition and a postposition that surround Arg2, an NP referring to an eventuality, (20).

(20) Um *der Verlängerung des Friedens* willen **sage ich jetzt nichts.**
'For the continuation of the peace, I am not saying anything.'

Prepositions are not annotated at all in the PDTB-2; some are being annotated in the PDTB-3 version, and in corresponding projects on other languages. It should be noted that it can be hard to distinguish between a discourse use and a non-discourse use of, e.g., an expression of the form *Prep + Vinf*. This is the case for *pour* in French (Colinet et al., 2014) or *to* in English, which introduces a "catenative" complement or a purpose adjunct: Huddleston and Pullum (2002, page 1223) point out that (21a) is ambiguous between a catenative interpretation glossed in (21b) and a purpose adjunct interpretation glossed in (21c).

(21) a. He swore to impress his mates.

b. = He swore that he would impress his mates.

c. = **He swore** in order to *impress his mates.*

Along the same lines, the annotation of expressions of the form *Prep + NP* requires determining whether the *NP* refers to an eventuality, which can be a difficult problem. This may well be, but prepositions and postpositions should nevertheless be included in discourse connective lexicons, even if they are not annotated in all corpora.

#### 3.1.3 COORDINATING CONJUNCTIONS

Each of the three languages we study here has a closed list of less than ten coordinating conjunctions (e.g., in French, *mais, ou, et, donc, or, car, ni*), although the status of some elements is not clear: for example, *puis* in French shares properties of both coordinating conjunctions and adverbs, and grammars differ on assigning its category.

Coordinating conjunctions basically introduce a finite clause (22a). However, the finite clause can be elliptical and reduced, for instance, to a VP, or to two NPs, when gapping (22b). They are positioned in front of the phrase they introduce, whose content is their Arg2, and they are placed to the right of their Arg1. They can never be externally modified.

(22) a. **John will help Mary** and *Ted will help Sue.*

b. **John will help Mary** and *Ted Sue.*

### 3.2 Inter-sentential primary connectives

Inter-sentential connectives are single-word adverbs *(next, finally, conversely)* or adverbial prepositional phrases (PPs) *(in summary, in conclusion, for example)*. Their host sentence is generally a finite clause. They appear at the beginning of or within Arg2, as in (23), and the order is (obligatorily) Arg1-Arg2 hence, inter-sentential connectives never appear at the very beginning of a discourse).

(23) a. **Fred lost his keys.** Therefore, *he is in a bad mood.*

b. **Fred lost his keys.** *He is* therefore *in a bad mood.*

As we pointed out earlier, primary connectives of the form PP are frozen and non-compositional expressions, in contrast to secondary connectives. As an illustration of this difference, consider French *à part ça*: in (24a), the pronoun *ça* does not have an anaphoric reading, and the connective is a primary one; on the other hand, in (24b), *ça* has an anaphoric reading and the connective is secondary.[8]

(24) a. **Fred vient de s'acheter un costume de marque.** A part ça, *il se plaint qu'il est fauché.*
'Fred just bought a branded suit. Yet, he complains he is broke.'

b. **Fred vient de rater un examen.** A part ça, *il est en pleine forme.*
'Fred just failed an exam. Apart from that, he is in a good mood.'

A similar contrast is observed with the expression *à ce moment là* with both a primary and secondary connective use, see (25) from (Roze et al., 2012).

(25) a. **Tu as l'air de penser qu'elle n'est pas honnête.** A ce moment là, *ne lui raconte rien.*
'You seem to think she's not honest. So don't tell her anything.'

b. **Il a commencé à pleuvoir.** A ce moment là, *Fred est arrivé.*
'It started raining. At that moment, Fred arrived.'

### 3.3 Pairs of primary connectives

In Section 2.2, we pointed out that 'redundancy' is not a very good criterion for an insertion test for the presence of an "alternative lexicalization" in a sentence. In the following, we elaborate on this point by discussing instances where two primary connectives form a pair that jointly contributes to discourse coherence, and we distinguish the two cases of the connectives being present in the same argument, or are split between Arg1 and Arg2.[9] We want to stress that pairs of connectives should not be confused with complex multi-word primary connectives such as *when and if* in English or *wenn .. auch* in German, which were described in Section 3.1.

#### 3.3.1 BOTH PARTS OF THE PAIR HOSTED BY THE SAME ARGUMENT

"Doublets of connectives" are made up of either two adverbial connectives or a subordinating conjunction and an adverbial connective. The two connectives in a doublet share the same two arguments and express roughly the same sense, as in (26). In a doublet, the first adverbial or the subordinating conjunction has a somewhat more general meaning than the second adverbial: the second adverbial can signal an additional facet of meaning, without fundamentally altering the coherence relation. In (26a), *then* signals temporal succession, *subsequently* adds the information that the temporal intervals of the two events meet; in (26b), *but* signals an adversative relation, *however*

8. The (non)-anaphoric reading of *ça* is determined by testing the replacement of *ça* with a noun phrase such as *le fait que Arg1* ('the fact that Arg1'): this test leads to an incoherent discourse in (24a) and a coherent (though infelicitous) discourse in (24b).
9. For a more extensive treatment of this issue of connective pairs in German, see Stede and Irsig (2012).

more specifically a concessive one; in the German example (26c), *weil* ('because') signals a causal relation, *nämlich* ('in fact') does the same and adds a facet of speaker involvement (Grabski, 2008). In (26d), both *sondern* and *vielmehr* together signal a corrective replacement.

(26) a. **Fred finished his home work.** Then subsequently *he went to the movie.*

b. **Fred is a nice guy** but *he may* however *be tough with women.*

c. **Ich fliege nach Berlin,** weil *dort* nämlich *ein Bruder von mir wohnt.*
('I'm flying to Berlin *because* a brother of mine is living there.')

d. **Das scheint kein Lama zu sein,** sondern *es sieht* vielmehr *aus wie ein Dromedar.*
('That doesn't seem to be a llama *but* it *rather* looks like a dromedar.')

The crucial point to note is that, in a doublet, both connectives signal by themselves the coherence relation expressed in the examples. Therefore, it is important not to confuse these doublets of connectives with cases of "multiple connectives". Multiple connectives are hosted by the same Arg2 but have different senses and possibly different Arg1s, see (27). In (27a), taken from (Forbes-Riley et al., 2016), the conjunction *because* has a causal meaning and its Arg1 is the canceling event, while the adverbial *then* has a temporal meaning and its Arg1 is the ordering event. In (27b), the conjunction *but* and the adverbial *next* share the same arguments; *but* has an adversative meaning, *next* is used to make a partition of the thesis between its beginning and its end, as analyzed in (Danlos, 2005).

(27) a. John ordered three cases of Barolo. But he had to cancel the order because then *he discovered he was broke.*[10]

b. **The thesis begins with a brillant state of the art**, but next *it describes uninteresting experiments.*

We want to emphasize that multiple connectives should not be recorded in lexicons: *a priori,* any two connectives (signalling different relations) can be hosted by the same Arg2 as long as syntactic constraints are respected. On the other hand, the above-mentioned doublets of connectives (signalling the same relation) cannot be arbitrarily combined and thus should be recorded in lexicons, as we will discuss in Section 3.5.

### 3.3.2 PARTS OF THE PAIR DISTRIBUTED ACROSS ARGUMENTS

"Parallel connectives" are "pairs of connectives where one part presupposes the presence of the other, and where both together take the same two arguments," (PDTB Group, 2008). They are illustrated in (28). It has to be noted that one of the parts can be optional, and the other be a 'simple' connective in its own right. This holds for the three examples in (28): *on the other hand*, *if* and *or* can be used without their counterparts. The meaning can change, however: *either .. or* marks a strictly exclusive disjunction, while *or* does not mark in-/exclusion.

(28) 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.

10. As Arg1 is not the same for *because* and *then*, no boldface is used in this example.

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

c. Either sign new long-term commitments to buy future episodes or risk losing "Cosb" to a competitor.

Parallel connectives can be recorded in lexicons, where the conditions on optionality need to be stated in some way (see Section 3.5). The notion of Arg1/Arg2 does not map straightforwardly to parallel connectives and requires a revised definition, but this issue is left aside here.[11]

### 3.4 Conclusions on primary connectives

Concerning the possible types of arguments, we have seen that in accordance with the syntactic category of a connective, its Arg2 can be a finite clause, a VP or an NP referring to an eventuality. Hence, the AdjP category has been left aside. However, there are some primary connectives that introduce an AdjP. For example, *although*, on top of introducing a finite clause (29a) and a gerund participial VP (29b), can introduce an adjectival phrase (29c), and the same situation holds for concessive connectives in French and German (16a).

(29) a. Although *he is ill,* **Ted went to work.**

b. Although *being ill,* **Ted went to work.**

c. Although *ill,* **Ted went to work.**

We have not included verbal suffixes in the list of intra-sentential connectives. Indeed, in languages such as Turkish and Japanese, some verbal suffixes are frequently used to connect two discourse segments and are considered as "subordinators"; see (Zeyrek and Webber, 2008) for Turkish. In English (resp. French) the verbal suffix *-ing* (resp. *-ant*) could be considered as a subordinator with a result sense in examples such as *Fred made a sex joke, shocking Mary* ('Fred a fait une blague porno, choquant Marie'). This is left aside for further research.

So, putting aside verbal suffixes, we have described three categories of intra-sentential primary connectives (subordinating and coordinating conjunctions, adpositions) and two categories of inter-sentential primary connectives (adverbs and adverbial PPs). This leads to the question whether these five syntactic categories cover all the primary connectives. We believe this is by and large the case. Of course, as said in (Prasad et al., 2010), there are some frozen expressions that do not fall in these categories (such as *what's more* or *never mind that*); but in languages such as French or German, for which our earlier work has aimed at developing exhaustive lexicons for primary connectives, we found that only 3.2% and 3.6% of their entries, respectively, do no fall into our five categories. These low proportions show that these expressions can be considered as exceptions to the rule of primary connectives belonging to a closed list of syntactic categories (for the languages we are concerned with).[12]

11. Parallel connectives are quite frequent in Chinese, and so Zhou and Xue (2012) identify the two arguments of parallel connectives on semantic grounds, without using the notation Arg1/Arg2.
12. This claim should be reinforced by corpus frequencies: it is clear that idiomatic expressions such as *what's more* or *never mind that* appear much less frequently in corpora than conjunctions such as *because* or PPs such as *for example*.

## 3.5 Building lexicons for primary connectives

Having discussed the linguistic "map"s of primary connectives, we now turn to the task of actually representing the relevant information in a lexicon that, ideally, is readable for both humans and machines. In this section, we explain the design of the two early lexicons DiMLex (Stede, 2002; Scheffler and Stede, 2016) and LexConn (Roze et al., 2012; Danlos et al., 2015), point to recent extensions, and provide recommendations for starting similar projects in other languages.

### 3.5.1 ACQUIRING THE SET OF LEXICAL ITEMS

The first step in building a lexicon is to decide *which* items to include, along the lines of the definitions we provided in Sections 3.1 and 3.2. DiMLex in its early stages was developed in cooperation with the research group at Institut für Deutsche Sprache (IDS) that was responsible for compiling the "Handbook of German Connectives" (Pasch et al., 2003; Breindl et al., 2015). The first DiMLex version was a subset of the words studied by that group, viz. the 175 items that we intuitively considered as most frequent in present-day German. The latest DiMLex described in (Scheffler and Stede, 2016) with 276 entries resulted from another comparison with the final list of the IDS publication, but we did not include a number of items that we saw as outdated, or where we did not agree that they should be treated as connectives. Further, the "Handbook," for theoretical reasons, excludes prepositions, while DiMLex includes many of them. Notice that when compiling an inventory of connectives, the notion of "entry" is not trivial to define; this will be discussed in the next subsection.

For LexConn, a first version was obtained by compiling a list of elements belonging to the syntactic categories described above, which were manually filtered; the result was a lexicon with 325 entries (Roze et al., 2012). An annotation project supplemented this lexicon with an additional 30 entries (Danlos et al., 2015). As the annotation enterprise dealt with 18,535 sentences with 10,429 connective tokens annotated, we believe that this latests version is now nearly exhaustive.

When the goal is to build a lexicon for a new language, the starting point is the assumption that primary connectives form a closed class; the sizes of DiMLex and LexConn can be seen as indicative for the target number. One way of collecting a first base set of connectives is to work through standard grammars of the target language, which give explanations on conjunctions, adpositions and adverbials. Grammars that provide semantic classifications (as, for example, many grammars for second language learning do) can be especially helpful, because they also give hints on meaning of connectives.

As a potential alternative, we suggest taking advantage of a *parallel corpus*, in conjunction with an existing lexicon in one of the languages of the parallel corpus. If the corpus covers the target language $T$ and if there exists a connective lexicon in the the source language $S$, the parallel corpus can be used to automatically retrieve potentially-corresponding words for the connectives known in $S$.[13] The prerequisite is that the text pairs be both sentence-aligned and word-aligned. Then the $T$ words that are aligned to the already known connectives in $S$ can be retrieved. This process yields of course only an approximation, as it will contain three types of errors:

- Wrong word alignment: A connective in $S$ may be falsely aligned to some word in $T$.

13. A popular resource is the multilingual *Europarl* corpus (http://www.statmt.org/europarl/); it is also included in the larger collection at http://opus.nlpl.eu/. *Europarl* was extensively used for work on connectives in the project COMTIS (http://www.idiap.ch/project/comtis).

- Ambiguity: The word in $S$ may be used in a non-connective reading, so that the (possibly correctly) aligned word in $T$ should not be part of the connective lexicon in $T$.
- Translation as non-connective: A connective in $S$ may be correctly aligned to a word in $T$ that is not a connective in the target language.

Furthermore, the set of connectives determined for $T$ in this way is obviously not likely to be complete. Nonetheless, this procedure can significantly speed up the process of "bootstrapping" a lexicon (and, by running the process also in the opposite direction, also to validate the source lexicon). For the case of mapping German to Italian connectives (and backwards), the process is explained by Bourgonje et al. (2017). Since the meaning of the corresponding connectives in the two texts can expected to be similar, the senses can also be mapped from the $S$ lexicon to the $T$ one, for a start.

Finally, given a source lexicon in $S$, one may of course use manual or automatic translation resources in order to obtain a first version of a lexicon for $T$. This method was applied as the first step in building an Italian version of DiMLex (Feltracco et al., 2016).

### 3.5.2 ORGANIZING THE INFORMATION IN LEXICAL ENTRIES

For organizing the candidate items in a lexicon, an initial decision is to define the mapping from connective words to the notion of an "entry" in the lexicon: what do we regard as two variants of a single connective (to be treated in the same entry), and what is to be analyzed as two different connectives? A clear case is that of mere spelling variants of the same word (as were, for example, created by the German spelling reform in the 1990s). As long as no syntactic or semantic differences can be observed, these are listed as orthographic variants in a single entry, both in DiMLex and in LexConn.

Sometimes, a word can have multiple syntactic roles, with each of them being a connective. In this case, the lexical entry has to provide more than one field for syntactic information. This happens, for instance, with the German *trotzdem*, which can be a subordinating conjunction ('although') and in some dialects also an adverbial ('anyway'); in English, *though* shows a similar ambiguity. This situation needs to be distinguished from a word that also has a non-connective reading. In DiMLex, this is described in a separate 'ambiguity' part of the entry, which states whether or not such a reading is available, and also gives examples of non-connective uses. Notice that some words conflate these ambiguities, such as English *before*, whose three connective variants have been shown above in example (12), and which, in addition, is not a connective when heading an object-denoting NP (e.g., *before 1996*).

Next, there is semantic ambiguity in case a connective can signal more than one coherence relation, i.e., have more than one sense. DiMLex chose to encode this hierarchically embedded within the syntax field of an entry, which thus can have more than one semantic field, corresponding to each sense. (If the two senses also coincide with some syntactic difference, there have to be different syntax fields, though.)

A further complication for the question of "one entry" versus "two entries" arises when the connective can have two syntactic roles, coinciding with an orthographic variant. This is a frequent phenomenon both in German (e.g., *dadurch* - adverbial; *dadurch, dass* - complementizer) and in French (e.g., *avant de ('before')* - preposition; *avant que ('before')* - subordinating conjunction). In English, a corresponding question is the relationship between *because* (subordinating conjunction) and *because of* (preposition). In these cases, DiMLex and LexConn opt for two separate entries.

On the syntactic side, DiMLex chose not go deeply into detail (much less so than the resource compiled by Pasch et al. (2003) does). It provides the basic category of the connective as well as information on the possible orderings of the arguments: Arg1 precedes or follows Arg2, or Arg2 can be embedded in Arg1. In LexConn, this information is not recorded explicitly, but (see our discussion in Section 3.1) it largely follows from the syntactic category. Within the syntactic description, DiMLex gives the PDTB-3 sense relations and the frequencies that have been observed in a corpus study, where 25 instances of use (taken from the large www.dwds.de corpus) have been analyzed for each connective (Scheffler and Stede, 2016). In general, when building a lexicon for a new language, such a step of thoroughly considering corpus instances is to be highly recommended; recall that LexConn, too, has been finalized only after an accompanying corpus annotation project.

### 3.5.3 TECHNICAL IMPLEMENTATION

A fundamental decision made originally for DiMLex was to use XML as the technical format, because this allows for straightforward mappings for example, by means of XSL scripts) from the base lexicon to various format variants. These include a reduced version to be used in the semi-automatic annotation tool *Conano* (Stede and Heintze, 2004), an HTML version for the human reader using a web browser, and different language-technological applications such as discourse parsers. In addition, defining mappings to common spreadsheet formats is not difficult.

As an illustration, Figure 1 shows an abbreviated form of the DiMLex entry for *vielmehr*, which is similar to English 'instead' or 'rather', but is used predominantly in contexts of correction.

We use this example to explain the remaining aspects of DiMLex entries. The orthography is classified as being a simplex (single word) or multi-word connective (more than one word); for the latter, we additionally state whether the two components are discontinuous or not. Each orthographic variant receives its own identifier (an extension of the identifier for the complete connective), which allows for possible cross-referencing from other parts of the entry.

Following the ambiguity information, which consists of two binary features (i.e., with values 0 or 1), a feature specifies whether the connective can be in the scope of focus particles (cf. the discussion of example (14) in Section 3.1, repeated here in (30)).

(30) a. Only because *he wanted to see the MOMA,* **Fred went to New York.**

b. #Only as he wanted to see the MOMA, Fred went to New York.

A set of two features states whether it can occur in so-called 'correlate' constructions, which are loose yet not totally-arbitrary collocations between adverbials and conjunctions that we described as "doublets" in Section 3.3.1 above. In the example (Fig. 1), *vielmehr* is specified as a possible correlate of the conjunction *sondern* (a corrective form of 'but'). This means that when the adverbial *vielmehr* occurs in a *sondern*-clause, most likely the two words collectively signal the identical instance of a Substitution relation (instead of two different relations). An example with these two connectives was shown in (26d) above.

The DiMLex XML format has proven to be quite compatible with approaches to lexicons in other languages. We converted the original format of LexConn also to DiMLex XML, and new lexicons for Italian (Feltracco et al., 2016) and Portugese (Mendes et al., 2018) have recently been constructed following the DiMLex format. Likewise, we mapped the list of English connectives annotated in the PDTB project to DiMLex format and extended it with some 50 entries from other

```
<entry id="k152" word="vielmehr">
   <orths>
     <orth type="cont" onr="k152o1">
       <part type="single">vielmehr</part>
     </orth>
   </orths>
   <ambiguity>
        <non_conn freq="4" anno_N="25">1</non_conn>
        <sem_ambiguity>0</sem_ambiguity>
   </ambiguity>
   <non_conn_reading>0</non_conn_reading>
   <focuspart>0</focuspart>
   <correlate>
     <is_correlate>1</is_correlate>
     <has_correlate>0</has_correlate>
     <correlatee> <corr>sondern</corr> </correlatee>
   </correlate>
   <syn>
     <cat>konnadv</cat>
     <ordering>
       <ante>0</ante>
       <post>1</post>
       <insert>0</insert>
     </ordering>
     <sem>
      <pdtb3_relation sense="substitution-arg2-as-subst" freq="20" anno_N="25"/>
     </sem>
   </syn>
 </entry>
```

Figure 1: DiMLex sample entry: 'vielmehr' (abridged)

resources (Das et al., 2018). All of these resources, in a "minimalized" version covering the basic syntactic information and sense relations, have recently been made available online as an interlinked lexical database.[14]

## 4. Secondary discourse connectives

Above, we have defined primary connectives as either single-word units or multi-word units, which are non compositional, internally non-modifiable and non-inflected, while secondary connectives are multi-word units which are compositional, modifiable and inflectable. However, there are some exceptions to these rules. For example, the Czech primary connective *kdyby* ('if') may be inflected, see the forms *kdybych*, *kdybys*, *kdyby*, *kdybychom*, *kdybyste*, *kdyby* meaning 'if I, if you, if he/she/it, if we, if you, if they'. Therefore, primary and secondary connectives should not be considered as two strictly separated classes of expressions but rather as a scale of expressions in a different degree of grammaticalization. This means that due to language changes and possible increasing grammaticalization, secondary connectives may become primary in the future.[15] We should add that some primary connectives in one language translate as secondary connectives in another language. For example, the German primary connective *dagegen* does not have (yet) any fully grammaticalized counterpart in Czech (the Czech equivalent is *proti tomu* consisting of a preposition *proti* and an anaphoric pronoun in dative case).

Due to grammaticalization of primary connectives, it is possible to describe their formal characteristics in a comprehensive way, as was done in Section 3. However, secondary connectives, which are not (yet) fully grammaticalized, exhibit a high degree of variation and have thus to some extent an open form. As a consequence, it is only possible to describe their syntactic structures, which leaves room for variation. We regard these structures as "lexically headed," appealing to the notion of *core unit*: the core unit of a secondary connective is defined as the lexical unit with the strongest meaning. By this definition, any anaphoric element, whose meaning does not stand by itself, is not a core unit. Examples are: in the secondary connective *for this reason*, the core unit is the noun *reason*; in *this precedes*, it is the verb *precede*; in *because of this* it is the complex preposition *because of*. Given its strong compositional meaning, the core unit signals which type of discourse relation the secondary connective expresses. For example, a secondary connective whose core unit is *reason* can express either that Arg1 is the reason of Arg2 — as in the template given in (31a) — or that Arg2 is the reason of Arg1 — as in (31b). Similarly, when the core unit is the verb *precede*, either (the event in) Arg1 precedes (the event in) Arg2, as in (32a), or Arg2 precedes Arg1, as in (32a).[16]

(31) a. Arg1. For this reason Arg2.

b. Arg1. The reason is Arg2.

14. http://www.connective-lex.info

15. This idea is supported by historical origin of present-day primary connectives that arose from similar structures (and parts of speech) like 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*.

16. Using the PDTB hierarchy of senses tags, the core unit *reason* signals a relation of type CONTINGENCY-Cause, either Reason or Result, the core unit *precede* a relation of type TEMPORAL-Synchronous, either Precedence or Succession. Given the strong meaning of the core unit, the annotation of discourse relations expressed by secondary connectives in discourse corpora is slightly easier than for (semantically more general) primary connectives. This is reflected in the inter-annotator agreement of discourse relations in Czech where the agreement on relations expressed by secondary connectives was slightly higher than by primary connectives: 0.82 vs. 0.77.

(32) a. Arg1. This precedes Arg2.

b. Arg1. This was preceded by Arg2.

We describe below the list of the most frequent structures for secondary connectives, underlying for each structure the syntactic category of its core unit. These structures are illustrated with English examples, but there exist equivalents in Czech, French or German (and possibly many other languages). As for primary connectives, we distinguish intra versus inter-sentential secondary connectives.

## 4.1 Intra-sentential secondary connectives

### 4.1.1 SECONDARY SUBORDINATING CONJUNCTIONS AND PREPOSITIONS

This class includes PPs that introduce a complement in the form of a clause, a VP or NP, see (33). They are modifiable (*in the vain hope*) and they may appear without any complement (*in this hope*). With a complement (referring to Arg2), they can be qualified as secondary subordinating conjunctions or prepositions.

(33) a. **People were trained** in the hope that *they would find jobs.* (= 15)

b. **He was studying** in the hope of *being admitted to an engineering college.*

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

As far as we know, there are no multi-word units that can be qualified as (secondary) coordinating conjunctions.

## 4.2 Inter-sentential secondary connectives

### 4.2.1 ADVERBIAL PREPOSITIONAL PHRASES

Secondary connectives in the form of prepositional phrases (PPs) are of two kinds: the first is a combination of a preposition and an anaphoric expression, mostly a demonstrative pronoun, like *due to this*, *because of this*, *despite this*, *besides this*, *thanks to this*, *in spite of this*, as in (34). The core units of these secondary connectives are the prepositions i.e., *due to*, *because of*, *despite*, which signal the semantic types of discourse relations e.g., *despite* signals a concession relation.

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

The second type of secondary connectives in the form of PPs consists of a preposition and a content noun which is the core unit, like *for this reason*, *under these conditions*, *for this purpose*, (35).

(35) **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. These schemes can be used in lexicons for secondary connectives, which will be discussed in Section 4.3.

As it is the case for adverbial primary PPs, Arg2 is syntactically expressed by a clause and appears obligatorily after Arg1. Secondary PPs appear most of the time at the beginning of Arg2, but it may happen that they appear within Arg2.

### 4.2.2 DISCOURSE VERBS

Other secondary connectives contain a semantically strong verb that forms the core unit, e.g., the verb *mean* (occurring in the secondary connective *it means that*) or other verbs like *cause*, *precede*, *follow*, *prove* which are called "discourse verbs" in (Danlos, 2006). The subject of a discourse verb in a secondary connective is an anaphoric pronoun referring to Arg1. The examples in (36) show that Arg2 can be nominal or clausal.

(36) a. **CL start to rise, reaching the maximum level, twice that of healthy controls, on day +11.** This preceded *the rise of blood leukocytes above 1.0X10(9)1.*

b. **Most researches in this field, on the way to understand challenges and propose solutions, are based on case studies.** This causes that *findings are confined to particular situations.*

Discourse verbs can be used in the active or passive form (e.g. *this was preceded by, this was caused by*). They are schematized as DVs.

### 4.2.3 COPULA STRUCTURES

Many secondary connectives contain a semantically weak verb (mostly *be*). First, this verb can be built with a subject whose head noun is the core unit like *reason*, *condition*, *consequence*, *example*, *conclusion*. The examples in (37) show that Arg2 can be nominal or clausal.

(37) 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 arrived at the capital.*

Another structure with the weak verb *be* is illustrated in (38), in which the core unit *reason* appears after the copula, the subject being an anaphoric pronoun referring to Arg1.

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

It should be stressed that Arg2 in (37) expresses the reason of Arg1 while Arg2 in (38) expresses the result of Arg1. These copula structures are schematized as BE/SubjN and BE/AttN respectively, which states that the core unit N appears as the subject or attribute of the copula.

The last structure with the weak verb *be* is lexically headed by a subordinating conjunction or a preposition, (39). The subject is an anaphoric pronoun referring to Arg1. The scheme is BE/Conj or BE/Prep.

(39) a. **Out in space, the sky looks black, instead of blue.** This is because *there is no atmosphere.*

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

#### 4.2.4 TO VINF PHRASES

Another frequent type of secondary connective is made up of the preposition *to* followed by an infinitival phrase, e.g. *to conclude, to sum up, to give an example, to make it short, to put it in a nutshell*. The core unit is either a plain verb (*conclude*) or a verbal expression (*give an example, put it in a nutshell*), (40). Arg2 is clausal.

(40) **Denotation and sense can be applied to a lexeme or a larger expression. The denotation and sense of a composite expression is ....** To put it in a nutshell, *denotation, reference and sense are closely related to one another.*

### 4.3 Building lexicons for secondary connectives

For secondary connectives, right now, only a Czech lexicon is under development, following annotation of the PDiT (Prague Discourse Treebank 2.0 (Rysová et al., 2016)); we are not aware of any other efforts. The format of such lexicons is therefore not yet as stable as it is for lexicons of primary connectives, which have been developed for several languages (see Section 3.5). Nevertheless, we think that the following sections give guidelines that will help the development of lexicons for secondary connectives in other languages. Section 4.3.1 briefly addresses the question of delimiting what can be a lexical entry for such a lexicon. Section 4.3.2 discusses how to organize the information under each entry.

#### 4.3.1 ORGANIZING THE INFORMATION IN LEXICAL ENTRIES

We have seen that there are possibly several different structures containing the same core unit, e.g., *for this reason*, *the reason is*, *that is the reason why* for the core unit *reason*. The question is whether some of them are only variants of the same secondary connective (and if so, which one is basic and could be presented as a representative for the others) or whether all of them are individual secondary connectives that should be considered separate lexical entries.

From the formal point of view, all these examples are individual (structurally different) expressions, and therefore a possible solution would be to treat them as individual lexical entries. However, in an alphabetical ordering of the lexicon, the connectives *for this reason* and *that is the reason why*, for example, would appear very far away from each other, which would be inconvenient for the (human) lexicon users. Moreover, this solution would lead to an enormous number of entries and an overall opacity of the lexicon. At the same time, we cannot consider structures like *for this reason* and *that is the reason why* simply to be variants of the same connective.

In conclusion, the solution may be to keep all the secondary connectives with the same core unit within a single entry (so that the user can easily find all structures containing the same core unit in one place) and at the same time to differentiate between distinct structures by simply listing them under the common core unit without any hierarchy. In other words, we consider core units to be "umbrella lemmas" of the individual structures containing them.

#### 4.3.2 DESCRIPTION OF SECONDARY CONNECTIVES WITH THE SAME CORE UNIT

The most common schemes for secondary connectives were described in the previous sections. An entry with a given core unit lists the schemes in which it appears. These schemes may need to be specified; for example, in the scheme PP/N, the preposition needs to be specified for each core unit N. In the lexical entry under the core unit *reason* given in (41), there are three schemes; for each

scheme, its name (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 core unit field; optionality is marked with parenthesis and alternatives are written within brackets. Under each scheme, the field Realizations gives concrete examples of the scheme.

(41) **Core unit: N = *reason***

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

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

As pointed out earlier, secondary connectives may be inflected (*for these reasons*). The question is how many variants should be recorded in the lexicon. On the one hand, it is useful to capture all substantial characteristics of connectives; on the other hand, the lexicon should not be overcrowded by details. So we suggest not to include all the variants in the lexicon, but just to include the information as to whether a scheme is inflectable and modifiable, through the binary features Inflection and Modification respectively (a typical example of a concrete inflected/modified realization can be added).

Note that the scheme in itself indicates where possible modifications can take place. For example, when the core unit is the noun *reason*, this noun can be modified by an adjective in the three schemes (41) where it appears (e.g., *for this simple reason, the main reason is, that is a possible reason why*). Moreover, when a scheme includes a verb (weak or plain), the verb itself can be modified (e.g., *the reason is possibly that, it will inevitably cause that, to conclude rapidly*). But this is a general phenomenon, so it should not be recorded in the lexicon. More generally, the scheme specifications (and the binary features) should lead to regular expressions that define search patterns to find secondary connectives in corpora.

The information given in (41) under the core unit *reason* 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 (42) for the first scheme under the core unit *reason*.

(42) **Core unit: N = *reason***

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

Inflection: 1
Modification: 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)
...

## 5. Semantics for primary and secondary connectives

So far, we have looked mainly at structural properties of primary and secondary connectives, while touching on semantics only occasionally. In this section, we consider first the question of the sense inventory for both groups, and then that of internal or external modifiers taking scope over a connective.

### 5.1 Senses of primary and secondary connectives

As said in Section 1, a discourse theory such as SDRT (Asher and Lascarides, 2003) and any discourse annotation project — e.g., (Prasad et al., 2008) for English or (Rysová et al., 2016) for Czech — relies on the assumption that discourse relations form a closed class of less than thirty elements (27 leaf nodes in the PDTB-3 hierarchy (Webber et al., 2016); 22 in Czech) that are considered to be the senses of primary connectives. In comparison, there exist around 300 primary connectives in French and German, and 150 in Czech. As a consequence, sense hierarchies do not attempt further semantic differentiation between primary connectives such as *therefore, thus, hence, consequently, as a result, as a consequence, as a side-effect*, which are said to all express the very same discourse relation (Result). Notwithstanding the coarse grain in these semantic analyses, there is a problem in that some connectives have an "open sense", i.e., a sense which does not fall into the closed list of senses. In French, 6.7% of primary connectives have an open sense. As an illustration, consider *au fur et à mesure que,* which expresses both temporal simultaneity and proportional equality between its arguments, as in (43a) whose accurate English translation is given in (43b) with no connective but a specific construction *the more ... the more*. The specific sense of this conjunction does not belong to the closed list, so it is considered in LexConn as an open sense. This choice means that a too coarse-grain semantic analysis is rejected: it is considered as inappropriate to give *au fur et à mesure que* just a simultaneity sense, which belongs to the closed list. We stress that the number of French primary connectives with an open sense is twice the number of French primary connectives with an "open syntactic category", i.e., a syntactic category that does not fall into the closed list presented in Section 3.

(43) a. **Son élocution devenait de plus en plus inaudible** au fur et à mesure qu'*il descendait la bouteille de whisky.*

b. The more he went down the bottle of whisky, the more inaudible his speech was.

What is the situation for the senses of secondary connectives? Following the position adopted for primary connectives, one can make the assumption that secondary connectives such as *the result is, the consequence is, a side-effect is, this results in, this causes that, this leads to, this brings about, for this reason* all express the very same Result relation. However, the number of secondary connectives with an open sense may be important, especially when considering those connectives whose lexical head is a noun with a compositional meaning. In Czech, we estimate that 15% of secondary connectives have an open sense (i.e., a sense which is not one of the 22 senses adopted for primary connectives). As an illustration, consider *v tomto ohledu* ('in this respect') which is illustrated in (44). It does not have any equivalent as a primary connective: it cannot be replaced, for example, by *tak* ('so') (with a Result sense) without changing the meaning of the discourse.

(44) *Rychlost internetu je v této zemi velmi malá.* V tomto ohledu **je daná země pro uživatele internetu jednou z nejméně výhodných zemí.**
'The internet speed is low in this country. In this respect, it is one of the worst countries to use the Internet in'.

One solution consists in adding new senses to the closed list, for example to add the sense 'Regard', as it has been adopted in the PDiT. This sense is expressed in Czech, English, French or German not by a primary connective but by various secondary connectives, for example in English by the following PPs (schematized as PP/N): *in this respect, in this regard, in this perspective, against this backdrop*.

However, the solution of adding new senses to the closed list is not appropriate for secondary connectives with a very specific sense such as *in the hope (that/of)*, which was illustrated in (33) in Section 4.1.1. When rejecting that its sense is Goal (in the closed list) because of the non-synonymy between *hope* and *goal*, it must be considered an open sense; it would be peculiar to add a sense 'Hope' just because of this connective.

In conclusion, while at the syntactic level, it seems clear that primary and secondary connectives belong to a closed list of syntactic categories/templates with only a few exceptions (Section 3.4), the situation is more confusing at the semantic level. Even though the list of senses for primary connectives appears to be fairly closed, as demonstrated by corpus annotation projects, this seems not to be the case for secondary connectives. Our strategy would be to introduce new senses if a variety of very similar secondary connectives exists in a language (preferably, in many languages), such as 'Regard', while leaving more idiosyncratic secondary connectives recorded in the lexicon with an open sense.

### 5.2 Modification of primary and secondary connectives

Another interesting aspect of connective semantics is modification. For primary connectives, only external modification is possible (Section 3.1) and modifiers are roughly focus particles (e.g., *mainly, only, just*) with a modal semantic value. For secondary connectives, internal modification is possible (Section 3.2) and internal modifiers can have a modal value (*a possible reason is, the main reason is, possibly this has caused*) or an evaluative value (*the awful consequence is, a positive result is, this was unfortunately followed by*). Moreover, two modifiers can be found within the same expression, as in the example *probably the most egregious example is*, as cited in (Prasad et al., 2010). Discourse theories essentially ignore modification of discourse relations with a modal value, let alone with an evaluative value. Discourse annotation faces difficulties with modification of connectives

and usually resorts to ad-hoc solutions. This means that research is needed on the modification of discourse relations/connectives (at the theoretical and practical levels) with solutions that work for primary connectives as well as for secondary connectives.

As an illustration, consider the authentic example (45). In discourse parsing or annotation, it can be computed or annotated that there is a Result relation between Arg1 (in bold) and Arg2 (in italics), due to the secondary connective *an undesirable side-effect is that* whose core unit is the noun *side-effect* used in a BE/SubjN structure (Section 4.2.3) and modified by the adjective *undesirable*. This analysis avoids placing an implicit connective between the two typographic sentences and can be good enough for many NLP tasks. However, it does not make any distinction between a result and a side-effect, and it does not state that the scope of the evaluative adjective *undesirable* is Arg2, which means that Arg2 is undesirable for the writer. This can be achieved in a deeper semantic analysis, which we here suggest as a goal for future research.

(45) **Conventional statistics-based methods for joint Chinese word segmentation and part-of-speech tagging have generalization ability to recognize new words that do not appear in the training data.** An undesirable side-effect is that *a number of meaningless words will be incorrectly created.* ... [CoRR 2013]

## 6. Conclusion and Outlook

In our proposal on structuring the realm of DRDs, we have provided clear definitions for primary and secondary connectives, showing that the former are often grammaticalized variants of the latter. The syntactic categories of primary connectives belong to a closed list (conjunctions, adpositions, adverbs, prepositional phrases), except for a very small number of other elements (and putting aside verbal suffixes). Exhaustive lexicons can be developed in a format that was first designed for German and more recently implemented for French, Italian, English, Portugese and other languages, so that we consider it as stable for this type of languages now. We explained the design choices and provided hints for creating a resource along these lines for a new language. One advantage of the XML format is that it allows for easily maintaining variants of a lexicon, which differ in granularity: for German, the smallest version includes only the information that is needed for the semi-automatic annotation tool Conano (i.e., syntactic category, example sentences, and sense), a slightly extended version is part of the parallel lexicon set available at `www.connective-lex.info`, and a yet larger version contains additional information that is still under construction and hence not published yet. Straightforward XSLT scripts can derive simpler versions from the "master" version.

Secondary connectives, which are compositional, modifiable and inflectable, appear in syntactic structures that are lexically headed by the core unit, i.e., the unit that has the strongest meaning. We have given a list of the most frequent structures and suggested guidelines to develop lexicons relying on these structures. So far, only a Czech lexicon is being developed for secondary connectives after annotation in a corpus, so we are not yet in a position to discuss coverage questions.

We hope that the joint work presented here will support the development of lexicons for primary and secondary connectives, as well as discourse annotation projects, in other languages. With a growing set of such resources, information on cross-language linking can be added, so that matters of translation can be studied. Also, in our future work, we plan to address various questions on semantics, e.g., to what extent the senses of connectives form a closed list, when multiple languages are being considered.

Finally, we wish to mention again the potential role of such lexicons for discourse parsers (either PDTB-style "shallow" or RST/SDRT-style "deep"), which can use a lexical resource to supplement the corpus-derived models of relation signals in order to improve recall on the task of identifying coherence relations. As long as the size of relation-annotated corpora remains relatively small — which is the case for any language but English today — such a lexicon can be of great help for bootstrapping approaches to parsing.

## Acknowledgments

We thank Ewan Dunbar for editing our English. We also acknowledge support from Institut Universitaire de France for the French part, the Czech Science Foundation project no. GA17-06123S for the Czech part, Deutsche Forschungsgemeinschaft project 'Anaphoricity in Connectives' and SFB 1287 for the German part, and the ISCH COST Action IS1312 'TextLink' for the collaborative part.

## References

Nicholas Asher. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht, 1993.

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

Peter Bourgonje, Yulia Grishina, and Manfred Stede. Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-IT 2017)*, Rome, 2017.

Chloé Braud and Pascal Denis. 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, 2016.

Eva Breindl, Anna Volodina, and Ulrich Herrmann Waßner. *Handbuch der deutschen Konnektoren 2*. Walter de Gruyter, Berlin/New York, 2015.

Margot Colinet, Laurence Danlos, Mathilde Dargnat, and Grégoire Winterstein. Emplois de la préposition *pour* suivie d'une infinitive : description, critères formels et annotation en corpus. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, and S. Prévost, editors, *4ème Congrès Mondial de Linguistique Française*, pages 3041 – 3058, Berlin, Germany, 2014.

Laurence Danlos. Partition of an entity with aspectuo-temporal operators. In *Proceedings of the Third International Workshop on Generative Approaches to the Lexicon (GL'2005)*, Genève, Switzerland, 2005.

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

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

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

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

Kerstin Fischer. *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles*. Mouton de Gruyter, Berlin/New York, 2000.

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

Kathy Forbes-Riley, Bonnie Webber, and Aravind Joshi. The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics 23(1)*, 2016.

Michael Grabski. Connectives that manage perspectives in discourse: On the function of German 'nämlich', 'schließlich', and 'also'. In *Proceedings of the 3rd Workshop on Constraints in Discourse (CID 2008)*, Potsdam, Germany, 2008.

Josef Hrbáček. *Nárys textové syntaxe spisovné češtiny*. Trizonia, Prague, Czechia, 1994. ISBN 80-85573-51-2.

Rodney Huddleston and Geoffrey Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, 2002.

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

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

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

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

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

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

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

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

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 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, 2010.

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

Magdaléna Rysová. *Diskurzní konektory v češtině (Od centra k periferii). [Discourse Connectives in Czech (From Centre to Periphery).] PhD Thesis.* Charles University in Prague, Prague, Czechia, 2015.

Magdaléna Rysová and Kateřina Rysová. 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, 2014.

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

Ted Sanders, Wilbert Spooren, and Leo Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35, 1992.

Tatjana Scheffler and Manfred Stede. 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, 2016.

Manfred Stede. 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, 2002.

Manfred Stede and Silvan Heintze. Machine-assisted rhetorical structure annotation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 425–431, Geneva, Switzerland, 2004.

Manfred Stede and Kristin Irsig. Complex connectives in German: Complications for local coherence analysis. In Anton Benz, Manfred Stede, and Peter Kühnlein, editors, *Constraints in Discourse 3 - Representing and Inferring Discourse Structure*, pages 165–182. Benjamins, Amsterdam, 2012.

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

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. A discourse-annotated corpus of conjoined VPs. In *Proceedings of the Tenth Linguistic Annotation Workshop (LAW-X 2016)*, Berlin, Germany, 2016.

Deniz Zeyrek and Bonnie Webber. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, Hyderabad, India, 2008.

Yuping Zhou and Nianwen Xue. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (COLING 2012)*, pages 69–77. Association for Computational Linguistics, 2012.

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