

Building discourse relational device lexicons

Laurence Danlos

Université Paris Diderot-Paris 7, Alpage, IUF, FRANCE

TextLink Training school, 21/01/2016

Outline of the talk

- ① Information in a discourse connective lexicon
- ② First Method to build such a lexicon: linguistic criteria
- ③ Second Method to build such a lexicon: corpus data
- ④ Third Method to build such a lexicon: cross-linguistic data

Outline of the talk

- 1 Information in a discourse connective lexicon
- 2 First method: linguistic criteria
- 3 Second method: corpus data
- 4 Third method: cross-linguistic data
- 5 Other related issues and Conclusion

Discourse connectives

- Discourse connectives form a subset of Discourse Relational Devices (DRDs).
- Discourse connectives: logico-semantic predicates with two arguments, named Arg_1 and Arg_2 , denoting eventualities.
- focus of the talk: linguistic criteria to identify discourse connectives in **written texts** (no speech or chat).
- Other DRDs: briefly sketched at the end of the talk.

An entry is a form of discourse connective with:

- its surfacic/orthographic variant(s) if any,
- its morpho-syntactic category (one entry by category),
- its sense (one entry by sense), generally through a set of discourse relations,
- example(s) of use,
- other miscellaneous information:
 - constraint on the position of the connective,
 - synonymy link with other connectives,
 - ambiguity with a non-discourse use,
 - etc

Morpho-syntactic categories of discourse connectives

Two major categories:

- “Intra-sentential connectives”
- “Inter-sentential connectives”

Linguistic criterion to distinguish these two major categories:

- **Intra-sentential** connectives form discourse segments that **can** be embedded under a matrix clause, (1)
- **Inter-sentential** connectives form discourse segments that **cannot** be embedded under a matrix clause, (2).

- (1)
- a. Fred is nice, **but** he may be tough with women.
 - b. Jane said that [Fred is nice, **but** he may be tough women].
 - c. Fred is nice. **But** he may be tough with women.
- (2)
- a. Fred is nice, **therefore** he is never tough with women.
 - b. *Jane said that [Fred is nice, therefore he is never tough with women].
 - c. Fred is nice. **Therefore**, he is never tough with women.

Intra- versus inter- sentential connectives

- These two categories are grouped under a single notion (“discourse connective”) only in the discourse community.
- In syntax or in formal semantics, these two categories are totally distinct and never studied together.
- In the rest of this talk, I will consider both intra- and inter-sentential connectives, even though their grouping under a single notion is debatable.

Categories within intra-sentential connectives: introduction

Coordinating conjunctions

- ex: *et (and)*, *mais (but)*, *ou (or)*
- they introduce a finite clause, so noted as *Coord + S*

Subordinating conjunctions

- ex: *si (if)*, *parce que (because)*, *pour que (in order that)*
- they introduce a finite clause, so noted as *Cunj + S*

Prepositions

- ex: *pour (in order to)*, *au lieu de (instead of)*
- they introduce a VP (possibly an NP, see later), so noted as *Prep + VP* (possibly *Prep + NP*)

Categories within inter-sentential connectives: introduction

- one-word adverbs: *ensuite* (next), *finalement* (finally), *inversement* (conversely)
- adverbial PPs (prepositional phrases): *en résumé* (in summary), *en conclusion* (in conclusion), *par exemple* (for example)
- their host is a finite clause, so noted *Adv + S*

Differences between intra- and inter- sentential connectives

Position of connectives

- Intra-sentential connectives (Coord/Cunj/Prep + X): compulsorily at the beginning of the element X they introduce, see (3)
- Inter-sentential connectives (Adv + S): at the beginning or within S, see (4).

- (3) a. Fred est de mauvaise humeur, **parce qu'** il a perdu ses clefs.
b. *Fred est de mauvaise humeur, il a **parce que** perdu ses clefs. (*Fred is in a bad mood, because he lost his keys.*)

- (4) a. Fred a perdu ses clefs. **De ce fait**, il est de mauvaise humeur.
b. Fred a perdu ses clefs. Il est, **de ce fait**, de mauvaise humeur.
(*Fred lost his keys. Therefore, he is in a bad mood/ He is, therefore, in a bad mood.*)

Differences between intra- and inter- sentential connectives ctd

Position of Conn + X

- Intra-sentential connectives (Cunj/Prep + X): Conn + X form a phrase that can be roughly anywhere within Arg_1 (except for Coord + S), see (5)
- Inter-sentential connectives (Adv + S): compulsorily after (on the right of) Arg_1 , see (6).

- (5) a. Fred est de mauvaise humeur, **parce qu'** il a perdu ses clefs.
b. Fred, **parce qu'** il a perdu ses clefs, est de mauvaise humeur. (*Fred, because he lost his keys, is in a bad mood.*)
- (6) a. Fred a perdu ses clefs. **De ce fait**, il est de mauvaise humeur.
b. *Fred, de ce fait il est de mauvaise humeur, a perdu ses clefs.
(*Fred, therefore he is in a bad mood, lost his keys.*)

Set of discourse relations (taxonomy)

Roughly the same as in the PDTB

Not a focus in this talk

LexConn: discourse connective lexicon for French

Version XML

Version web

Outline of the talk

- 1 Information in a discourse connective lexicon
- 2 **First method: linguistic criteria**
- 3 Second method: corpus data
- 4 Third method: cross-linguistic data
- 5 Other related issues and Conclusion

Building a lexicon with linguistic criteria

Two operations:

- Get a list of possible candidates (for example, from electronic dictionaries or from the translation of a connective lexicon in another language, see Third Method).
- Establish a set of linguistic criteria to filter this list:
 - The criteria may change according to the morpho-syntactic categories.
 - Comparison between French LexConn and German DiMLex.

Criteria for intra-sentential connectives: conjunctions

coordinating conjunctions *Coord + S*

in LexConn: all of them (*mais, ou, et, donc, or, car, ni*)

subordinating conjunctions *Cunj + S*

in LexConn: any unit which introduces a finite clause (adverbial clause) with or without a complementizer C:

- one word: *si (if), quand (when), comme (as)* ...
- Prep + C: *pour que (in order that), afin que (so that)* ...
- PP + C: *à condition que (on condition that), dans le but que (in the aim that), dans l'espoir que (with the hope that), au moment où (at the time when)*

in DiMLex, some PPs are excluded (see later)

Prepositions which introduce an infinitival or gerund-participial VP (*Prep + VP*)

in LexConn : any unit

- one word: *pour* (*in order to*), *après* (*after*), *sans* (*without*), *en* (*by*) ...
- ADV + C: *avant de* (*before*), *au lieu de* (*instead of*) ...
- PP + C: *dans le but de* (*in the aim of*), *dans l'espoir de* (*with the hope of*) ...

in DiMLex, some PPs are excluded (see later)

Prepositions which introduce an NP referring to an event (e.g. a nominalization)

in LexConn: no unit which introduces only an event NP

⇒ *grâce à* (*thanks to*), *à cause de* (*because of*) are excluded

in DiMLex: units which introduce only an event NP

⇒ *dank* (*thanks to*), *wegen* (*because of*) are included

Criteria for inter-sentential connectives: adverbs and PPs

No use at the beginning of a discourse

- (7) #*De ce fait*, Jean est de mauvaise humeur.
#*Therefore*, John is in a bad mood.

No use in a it-cleft clause

- (8) a. Il a joué dans plein de films. *Par exemple*, il tient le rôle principal dans Furo. *He played in a lot of films. For example, it takes the lead role in Furo.*
- b. #Il a joué dans plein de films. C'est *par exemple* qu'il tient le rôle principal dans Furo. *#He played in a lot of films. It is for example that it takes the lead role in Furo.*

Criteria for inter-sentential connectives: PPs

The basic idea is that these PPs must be non compositional and with a high degree of frozenness (grammaticalization process).

Compositional PPs could be considered as **AltLex** (alternative lexical unit, a notion introduced and explained in the PDTB).

Examples of **connective** in (9a) and **AltLex** in (9b):

- (9) a. Fred est de mauvaise humeur **parce qu'**il a perdu ses clefs.
*Fred is in a bad mood **because** he lost his keys.*
- b. Fred est de mauvaise humeur. **Ceci est dû au fait qu'** il a perdu ses clefs
*Fred is in a bad mood. **This is due to the fact that** he lost his keys.*

Examples in French : *après ça*, *à part ça*, *à ce moment-là*

Criterion : if the pronominal form is anaphorical, then it is compositional and excluded (10) ; otherwise it is kept (11).

- (10) a. Fred est allé en Argentine. **Après ça**, il est allé au Pérou.
b. Fred est allé en Argentine. Après qu'il est allé en Argentine, il est allé au Pérou. *Fred went to Argentina. After (that/he went to Argentina), he went to Perou.*
- (11) a. J'ai croisé Fred dans un nightclub. **À part ça**, il nous dit qu'il est fatigué.
b. #J'ai croisé Fred dans un nightclub. À part que je l'ai croisé dans un nightclub, il nous dit qu'il est fatigué. *I met Fred in a nightclub. Except (that/#I met him in a nightclub), he tells us he is tired.*

PPs: ambiguity between an Altlex and a connective

- (12) a. Il commença à pleuvoir. **A ce moment-là**, Fred arriva.
b. Il commença à pleuvoir. C'est **à ce moment-là** que Fred arriva.
c. Il commença à pleuvoir. **A ce moment-là précis**, Fred arriva.
It started raining. At this (very) moment, Fred arrived.

- (13) a. Tu penses qu'elle n'est pas honnête. **A ce moment-là**, tu devrais ne rien lui raconter.
You think she is not honest. Then, you should tell her nothing.
b. #Tu penses qu'elle n'est pas honnête. C'est à ce moment-là que tu devrais ne rien lui raconter.
c. #Tu penses qu'elle n'est pas honnête. A ce moment-là précis, tu devrais ne rien lui raconter.

PPs with a high degree of frozenness

The PP connective *à la place* (*instead*) in (14a) can have no variant in number (14b) and no modification (14c):

- (14) a. Fred n'est pas allé au cinéma. **A la place**, il a fait du jogging.
Fred didn't go to the movies. Instead, he went jogging.
- b. Fred n'est pas allé au cinéma. *Aux places, il a fait du jogging.
- c. Fred n'est pas allé au cinéma. *A la place précise, il a fait du jogging.

Prepositions which introduce an event NP

- NO in LexConn for French
- YES in DiMLex for German
- A decision could be made within TextLink

Pros and Cons

- it is easier to exclude them: (i) no need to disambiguate between NPs referring to an event and other NPs, (ii) less connectives to be annotated
- it is more coherent to include them: see (15)

(15) a. Fred est parti **après que** Marie a pris sa douche.

b. Fred est parti **après** avoir pris sa douche.

c. Fred est parti après sa douche.

Fred left after (Mary took a shower/taking a shower/his shower).

Non frozen intra-sentential connectives (16)

YES in LexConn, NO in DiMLex

Pros and Cons

- they should be **included** since they are synonymous with one-word intra-sentential connectives, (17)
- they should be **excluded** since multi-word inter-sentential connectives must be frozen

(16) Fred a fait une pizza, dans ce but/ dans un but précis/**dans le but de** faire plaisir à Marie.

Fred made a pizza, with this aim/ with a clear aim/ with the aim to please Mary.

(17) a. Fred a fait une pizza, **dans le but de/pour** plaire à Marie.

b. Fred machte eine Pizza, zum dem Zweck/**um** Marie bitte.

Fred made a pizza, with the aim/ in order to please Mary.

Inter-sentential PPs which include an anaphoric element

- NO in LexConn for French, (18)
- YES in DiMLex for German, (19)
- Idiosyncrasy for German?

- (18) a. Il commença à pleuvoir. **Après**, Fred arriva
b. Il commença à pleuvoir. **Après cela**, Fred arriva.
It started raining. After (that), Fred arrived.
- (19) a. Es begann zu regnen. **Wonach** Fred eingetroffen.
b. Es begann zu regnen. *Wo Nach /Nach Fred eingetroffen.

Assigning a sense to discourse connectives

- ① fix a set (taxonomy) of discourse relations \mathcal{S}
- ② find examples illustrating a given connective (for LexConn, examples from FranText)
- ③ assign to the given connective one or several sense(s) in \mathcal{S} if any
- ④ otherwise, use the sense MISSING (25 connectives in LexConn, 7%), see (20)

(20) a. Son élocution vacillait **au fur et à mesure qu'**il descendait la bouteille de whisky.

The more he was drinking the whisky bottle, the more his speech was flickering.

b. **Pour une fois qu'**il part à l'heure, il y a des problèmes dans le métro.

For once he leaves on time, there are problems in the subway.

Conclusion on building a connective lexicon with linguistic criteria

- The resulting lexicon is quite coherent on the linguistic level.
- It may be more or less complete:
 - missing connectives,
 - above all, for a given connective, missing sense(s)
- It may be improved with the second and third methods

Outline of the talk

- 1 Information in a discourse connective lexicon
- 2 First method: linguistic criteria
- 3 Second method: corpus data**
- 4 Third method: cross-linguistic data
- 5 Other related issues and Conclusion

Building a connective lexicon from a corpus

First idea : start from scratch

- browse a corpus and stop when a possible connective is found,
- assign a sense to this discourse connective and add it in the lexicon,
- keep on going.

Drawbacks

- Quite time consuming because a large corpus is needed to build a complete lexicon.
- The resulting lexicon could be incoherent on the linguistic level, for example with frozen and non frozen elements.

Second idea: start from an existing lexicon

FDTB project (French Discourse TreeBank)

- Discourse annotation on the corpus FTB, already syntactically annotated.
- Discourse annotation in the PDTB style.
- First step: FDTB1.

FDTB1 project (Danlos et al. 2015)

- Identification of **all** the occurrences of connectives in the corpus.
- Not only the first step in the complete discourse annotation,
- but also a way to improve LexConn.

- ① Project all the elements of LexConn on the corpus, i.e. all the occurrences of LexConn entries are highlighted.
- ② Filter out the occurrences which don't follow the LexConn linguistic criteria; step achieved thanks to the syntactic analysis, (21).
- ③ Disambiguate the remaining occurrences: 3 tasks.

(21) a. Fred made a pizza **pour** faire plaisir à Marie.

b. Fred made a pizza ~~**pour**~~ le plaisir de Marie.

c. Fred made a pizza ~~**pour**~~ Marie.

Fred made a pizza in order to please Mary/ for Mary's pleasure / for Mary.

First disambiguation task: morpho-syntactic disamb

- The word *car* is either a coordinating conjunction or a noun, (22)
- The sequence *en fait* is either an adverbial connective or a pronoun + verb, (23)

(22) a. Fred est de mauvaise humeur **car** il a perdu ses clefs
(*Fred is in a bad mood for he lost his keys.*)

b. Le ~~car~~ Paris-Pau arriva en retard.
(*The Paris-Pau bus arrived late.*)

(23) a. Fred a l'air heureux. **En fait**, il est gravement malade.
(*Fred looks happy. In fact, he is badly sick.*)

b. Cette place est piétonne. Le Maire ~~en fait~~ un parking.
(*This square is pedestrian. The Mayor makes it a parking.*)

Second disambiguation task for intra-sentential connectives

The subordinating conjunction *pour que* can be a connective but has also non-discourse uses, e.g. it can be subcategorized by a verb, a noun an adjective or an adverb, (24)

- (24) a. Fred a fait une pizza ~~pour que~~ Mary le félicite.
(Fred made a pizza in order that Mary congratulates him.)
- b. Fred va s'arranger ~~pour que~~ Marie garde les enfants ce soir.
(Fred will arrange for Marie keeps kids tonight.)
- c. Un ordinateur est nécessaire ~~pour que~~ j'accomplisse cette tâche.
(A computer is necessary for me to accomplish this task.)
- d. Il pleut trop ~~pour que~~ nous puissions faire une promenade.
(It rains too much for us to take a walk.)

Second disambiguation task for intra-sentential connectives

The preposition *pour* + VP-inf can be a connective but has also non-discourse uses, e.g. it can be subcategorized by a verb, a noun, an adjective or an adverb (25) (Colinet et al. 2014)

- (25) a. Fred a été puni **pour** avoir dit des gros mots.
(*Fred was punished things for saying bad words*)
- b. Le Maire n'a pas profité de l'occasion ~~**pour**~~ trancher.
(*The Mayor has not taken the opportunity to decide.*)
- c. Fred a trouvé une astuce ~~**pour**~~ peler les tomates.
(*Fred found a trick to peel tomatoes.*)
- d. Ce couteau est idéal ~~**pour**~~ peler les tomates.
(*This knife is ideal for peeling tomatoes.*)

Third disambiguation task for inter-sentential connectives

The adverb *ainsi* can be a connective but has also non-discourse uses, (26)

- (26) a. Fred a fait beaucoup de bêtises. **Ainsi** il a renversé sa tasse de café. (*Fred did a lot of stupid things. As an example, he spilled his coffee.*)
- b. Fred a ~~ainsi~~ dit au juge : “Va au diable”. (*Fred said to the judge: “Go to hell”.*)
- c. Fred se comporte ~~ainsi~~ quand il est fatigué. (*Fred behave this way when he is tired.*)

Third disambiguation task for inter-sentential connectives

The PP *d'abord* can be a connective but has also non-discourse uses, (27)

- (27) a. Fred a **d'abord** été au Pérou. Ensuite, il est allé au Chili.
(*Fred first went to Peru. Next, he went to Chili.*)
- b. Manger bio est ~~d'abord~~ une question de bon sens.
(*Eating organic is primarily a matter of common sense.*)

Quantitative data on FDTB1

FTB		FDTB1	
articles	1005	adverbials	3221
sentences	18535	coord conj	3653
words	535 000	sub conj	1949
		prép V-inf	1070
		en V-ant	536
		TOTAL	10429

Enhancement of LexConn thanks to FDTB1

- 5 connectives have been suppressed from LexConn.V1,
- 30 connectives have been added,
- In total, LexConn.V2 has 353 entries,
- The sense(s) of the connectives have been refined thanks to corpus examples.

Results on uses of LexConn connectives in a corpus

- Nearly 70% of elements in LexConn has an occurrence in the corpus FTB
- List of the 100 elements which are the most frequent
- List of the LexConn elements which are morpho-syntactically ambiguous with non-discourse use examples
- List of 3 subordinating conjunctions and 5 prepositions + VP-inf which have a non-discourse use with examples
- List of 100 inter-sentential connectives which have a non-discourse use with examples
- List of 50 inter-sentential connectives which have always a discourse use in the FTB

Outline of the talk

- 1 Information in a discourse connective lexicon
- 2 First method: linguistic criteria
- 3 Second method: corpus data
- 4 Third method: cross-linguistic data
- 5 Other related issues and Conclusion

Third method: cross-linguistic data

To build a new lexicon or to enhance an existing one

TextLink Short-term scientific mission of M. Colinet in Postdam (Colinet et al. 2016)

- translate the elements of LexConn in German and compare with DiMLex
- vice-versa, translate the elements of DiMLex in French and compare with LexConn

Warning

A translation correspondence between two units $unit_a \rightarrow unit_b$ from Lex_a and Lex_b doesn't mean that **any** occurrence of $unit_a$ in a corpus \mathcal{C}_a translates as $unit_b$ in an aligned corpus \mathcal{C}_b (Cartoni 2014).

Constraints for the list of elements in the two lexicons

As usual in translation, a one-word unit may translate as a multi-word unit or vice-versa, without changing the morpho-syntactic category,
e.g. *au lieu de* -> *statt*, both *Prep + VP*, (28)

- (28) a. Il est allé à Bruxelles, **au lieu d'**aller à Paris.
b. Er ist nach Bruxelles gegangen, **statt** nach Paris zu gehen.
(*He went to Bruxelles, instead of going to Paris*).

Should the morpho-syntactic category of $\text{unit}_a \rightarrow \text{unit}_b$ be the same?

- The two major classes of unit_a and unit_b — intra- and inter-sentential connectives — should be the same,
- which means, for example, that an adverb (inter-sentential connective) in Lex_a cannot translate as a conjunction (intra-sentential connective) in Lex_b or vice-versa.
- arguments: these two classes of connectives are too different on several linguistic grounds, they are not “comparable”

Should the morpho-syntactic category be the same within intra-sentential connectives?

Categories: *Coord + S*, *Sub + S*, *Prep + VP*, *Prep + NP*

The linguistic criteria should be agreed upon, e.g. if *Prep + NP* are excluded (resp. included) in Lex_a , then they should be excluded (resp. included) in Lex_b

Discrepancies could be found and allowed
e.g. *Prep + VP* \rightarrow *Sub + S* (29)

- (29) a. Marie prend une douche **avant d'** aller au lit.
b. Marie nimmt eine Dusche **bevor** sie ins Bett geht.
(*Mary takes a shower before going to bed.*)

Should the morpho-syntactic category be the same within inter-sentential connectives?

Categories: *adverbs + S* and *PPs + S*

only difference: one-word adverb versus multi-word PPs

Discrepancies can be found and allowed,

e.g. adverb -> PP, see *conversely* -> *à l'inverse*

Problems with the frontier between connectives and AltLex

French *en plus* (connective) and *en plus de cela* (AltLex) both translate in German as *dazu* (in which *da* is a pronominal form)

Constraints on the senses of connectives in Lex_a and Lex_b

By definition

a translation correspondence between two units $\text{unit}_a \rightarrow \text{unit}_b$ means that these two units have the same sense

As a consequence

the lexicon Lex_a may benefit from the senses recorded in Lex_b and vice-versa (Stede 2016)

Outline of the talk

- 1 Information in a discourse connective lexicon
- 2 First method: linguistic criteria
- 3 Second method: corpus data
- 4 Third method: cross-linguistic data
- 5 Other related issues and Conclusion

Other related issues and conclusion

- Other issues around the notion of connectives
- DRDs which are not discourse markers

Pairs of connectives

si ... alors, d'abord ... ensuite, d'une part ... d'autre part, (30)

- (30) a. **Si** il ne pleut pas, **alors** Fred sera heureux.
(If it doesn't rain, then Fred will be happy.)
- b. Fred a beaucoup voyagé cet été. **D'abord**, il est allé au Pérou. **Ensuite**, il est allé au Chili.
(Fred travelled a lot last summer. First, he went to Peru. Next, he went to Chili.)
- c. Fred est vraiment odieux. **D'une part**, il est radin. **D'autre part**, il est dur avec les femmes.
(Fred is really nasty. On the one hand, he is stingy. On the other hand, he is tough with women.)

Doublets of connectives

- two connectives with the same host sentence, the same sense and the same Arg1
- redundancy : *mais cependant, et puis*, (31)

- (31) a. Fred est généreux **mais cependant** il est radin.
(*Fred is generous but however he is stingy.*)
- b. Fred est allé au Pérou. **Et**, il est **ensuite** allé au Chili.
(*Fred went to Peru. And, he went next to Chili.*)

Meta-expressions

They look as connectives but they have only one argument and behave as speaker-oriented adverbs (32)

- (32) a. Pour parler franchement, Fred est idiot.
(*Frankly speaking, Fred is an idiot.*)
- b. Franchement, Fred est idiot.
(*Frankly, Fred is an idiot.*)

- Discourse connectives form a half-open (or half-closed) class, with more than 100 elements and less than 500 for a given language.
- Quite possible to build a discourse connective lexicon,
- with well established linguistic criteria,
- with enhancement from corpus annotations and/or from cross-linguistic data.

Goal in TextLink:

Build as many discourse connective lexicons as possible!