



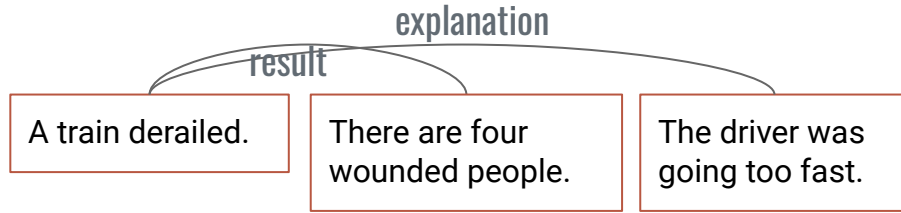

**KEEP
CALM
AND
ANDIAMO
AVANTI**

AnDiAMO: Analyzing Discourse Automatically, with Multiple Objectives

— Séminaire CLLE - 13/11/2023 —

Chloé Braud - CNRS - IRIT (Toulouse) - chloe.braud@irit.fr

Travaux en collaboration avec **Philippe Muller**, **Eleni Metheniti**, **Laura Rivière**, Marita Hernadez, Khushnur Binte Jahangir, Amir Zeldes, Janet Liu, Chuyuan Li, Nicolas Devatine, Mathieu Morey, Charlotte Roze, Pascal Denis, Mikel Iruskieta, Maximin Coavoux, Anders Sogaard et Ophélie Lacroix.



"Why did the train derail?"

*Whenever we read something closely, with even a bit of sensitivity, **text structure leaps off the page at us**. We begin to see elaborations, explanations, parallelisms, contrasts, temporal sequencing, and so on. These relations bind contiguous segments of text into a **global structure for the text as a whole**.*

Plan de la présentation

1. L'analyse discursive, en théorie et en pratique
2. ANR AnDiAMO : objectifs
3. Segmentation discursive
4. Relations discursives
5. Parsing discursif
6. Applications

→ Approches : gérer le manque de données

→ Questions :

- type d'information linguistique nécessaire pour une tâche discursive
- similarités entre les langues / modalités pour ce niveau d'analyse
- lien avec les cadres théoriques
- utilité de l'analyse discursive en TAL

Analyse discursive en TAL

- Cohérence et cohésion
- Relations discursives
- Cadres théoriques / schémas d'annotation
- Systèmes automatiques

Analyse discursive = dépasser le cadre de la phrase

Phénomènes linguistiques qui couvrent plusieurs phrases :

- Topics (topic segmentation),
- Liens temporels
- Anaphores et co-référence
- Relations discursives / rhétoriques

Structure discursive :

- révéler la cohérence textuelle
- interpréter des documents (i.e. inférences sur le contenu)

Liens entre les formes d'organisation textuelle, e.g. :

- contraintes discours / temporel
 - e.g. souvent l'effet après la cause
- discours / topic
 - e.g. certaines relations doivent rester sur le même topic
- discours / coréférence
 - e.g. certaines relations bloquent des référents potentiels

Cohérence et cohésion

- Document: pas une séquence aléatoire de phrases
 - A text is **cohesive** if its elements are linked together (non structural textual relations)
 - A text is **coherent** if it makes sense (structural relation between segments).

→ Document = un groupe de phrases cohérent et structuré

Exemples:

- 1) Paul fell. Marie pushed him. (Cause)
- 2) Paul fell. He likes spinach. (??)
- 3) Paul went to Istanbul. He has to attend a conference. (Reason)
- 4) Paul went to Istanbul. He likes spinach. (? Reason)

Cohérence et cohésion

Cohésion mais pas de cohérence:

i.e.: Chaque phrase est liée à celle qui précède, via des éléments lexicaux et grammaticaux, mais le texte n'a pas de sens

(1) I am a teacher. The teacher was late for class. Class rhymes with grass. The grass is always greener on the other side of the fence. But **it** wasn't. (Teacher resource site)

Résumé automatique: ni cohésif ni cohérent:

i.e.: mauvais ordre des phrases, pronoms sans antécédents

(2) It's like going to disney world for car buyers. I have to say that Carmax rocks. We bought **it** at Carmax, and I continue to have nothing bad to say about that company. After our last big car milestone, we've had an odyssey with cars. [Mithun and Kosseim, 2011]

Relations discursives

Les connecteurs discursifs sont des indices de l'existence des relations :

- contribuent à la cohésion et à la cohérence
- explicitent la relation entre les unités de texte adjacentes
 - **cependant** : signale une relation contrastive
 - **de plus** : indique que le texte qui suit élabore ou renforce le point présenté juste avant,
 - **pendant ce temps** : indique que deux événements sont contemporains
 - **si...alors** : indique une relation conditionnelle.

[Lexiques de connecteurs](#) disponibles pour différentes langues

Relations discursives

Les relations discursives lient :

- le contenu sémantique de deux unités (1)
 - ou un acte de parole et le contenu sémantique d'une autre unité (2)
 - on retrouve ces relations sans la présence de connecteur = implicite (2)
 - on retrouve ces phénomènes entre phrases et à l'intérieur des phrases (3)
- 1) This cute child turns out to be a blessing and a curse. She gives the Artist a sense of purpose, **but** also alerts him to the serious inadequacy of his vagrant life. (**Cause-reason**)
 - 2) Mrs. Yeargin is lying. They found students (..) who said she gave them similar help. (**Pragmatic Cause-justification**)
 - 3) Typically, money-fund yields beat comparable short-term investments **because** portfolio managers can vary maturities and go after highest rates. (**Cause-reason**)

Cadres théoriques : un consensus partiel

- Niveau d'analyse : le document
- Elementary Discourse Unit (EDU): clauses (en général), au plus une phrase
- Relations discursives : sémantico-pragmatiques, binaires (en général) et inter- ou intra-phrastiques
 - **Explicite** : avec un connecteur discursif, e.g. *Concession* en (1)
 - **Implicite** : absence de connecteur discursif, eg *Result and Sequence* en (2)

- 1) La chouette hulotte est un animal nocturne, **mais** elle peut vivre le jour.
- 2) Les tours se sont effondrées moins de deux heures plus tard (**Result**) **entraînant** l'immeuble du Marriott World Trade Center dans leur chute. (**Sequence**) La tour 7 du WTC s'est effondrée dans l'après-midi **en raison d'**incendies et des dégâts occasionnés par la chute des Twin Towers.

(Annodis)

Cadres théoriques : un consensus partiel

- Niveau d'analyse : le document
- Elementary Discourse Unit (EDU): clauses (en général), au plus une phrase
- Relations discursives : sémantico-pragmatiques, binaires (en général) et inter- ou intra-phrastiques
 - **Explicite** : avec un connecteur discursif, e.g. *Concession* en (1)
 - **Implicite** : absence de connecteur discursif, eg *Result and Sequence* en (2)
 - mais que fait-on de *because of* en (2)? et *dragging* in (2)?

- 1) La chouette hulotte est un animal nocturne, **mais** elle peut vivre le jour.
- 2) Les tours se sont effondrées moins de deux heures plus tard (**Result**) **entraînant** l'immeuble du Marriott World Trade Center dans leur chute. (**Sequence**) La tour 7 du WTC s'est effondrée dans l'après-midi **en raison d'**incendies et des dégâts occasionnés par la chute des Twin Towers.

(Annodis)

Deux 'vues' sur la structure discursive

Structure discursive hiérarchique (RST, SDRT, DLTAG, GraphBank...)

- Structure : arbres/graphes couvrant les documents
- Tente de donner une interprétation aux documents
- Annotation difficile !

Cohérence locale (PDTB)

- "theory neutral": fondé sur le lexique
- Structure plate, pas de couverture totale des documents
- Accords inter-annotateurs plus élevés (corpus plus larges)

Structure discursive hiérarchique

Rhetorical Structure Theory [Mann and Thompson, 1988]

- **Trees**
- Relations: **author intentions**
- “more important” segment = nucleus (vs satellite)
- 1 relation per pair of segments

RST DT (English) [Carlson et al 2001]

- **78 relations, 16 classes**
- **385 documents** - Wall Street Journal articles / PTB
- The most used: trees, full coverage
- But old, mono domain, weird choices

GUM (English) [Zeldes]

- Multi-domain incl. interviews, News, travel guides, how-to guides, academic writing, biographies, ...
- 148 documents (in progress)

Segmented Discourse Representation Theory

[Asher and Lascarides, 2003]

- **Graphs**
- Relations: semantics
- Subordinating / coordinating relations
- Multiple relations between segments

Annodis (French) [Afantenos et al. 2012]

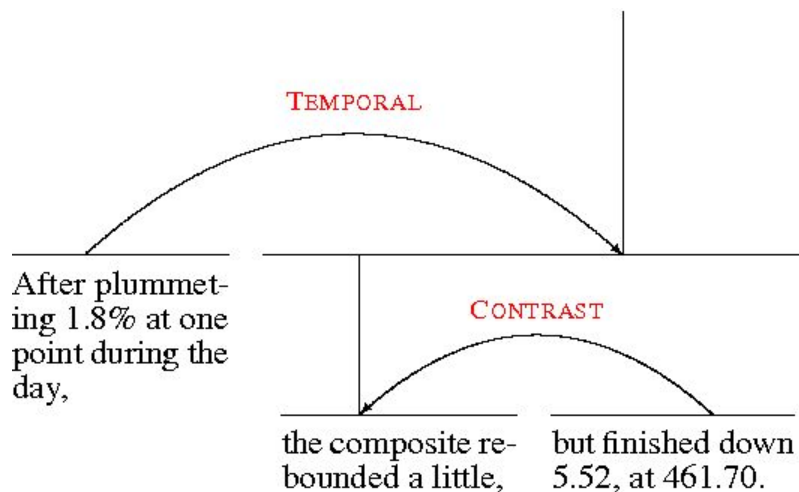
- **18 relations,**
- **86 documents** - Est Républicain + Wikipédia

STAC (English) [Asher et al. 2016]

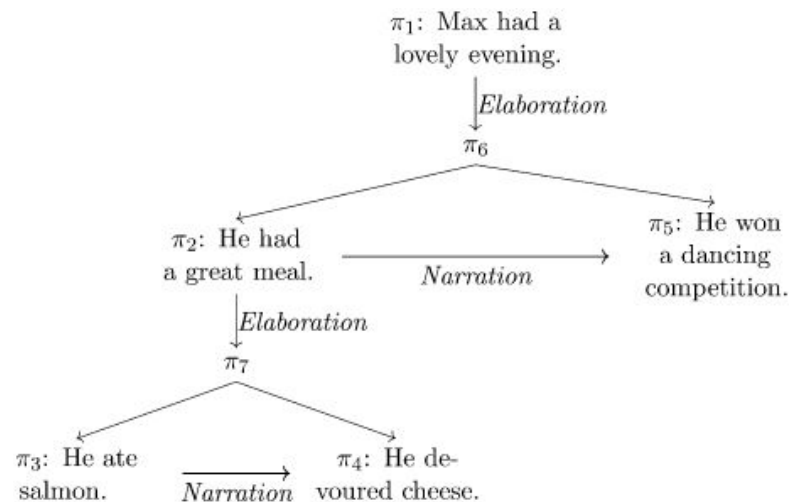
- 45 games, 1,091 (strategic) conversations
- multi-party chat dialogues

Structure discursive hiérarchique

RST tree



SDRT Graph



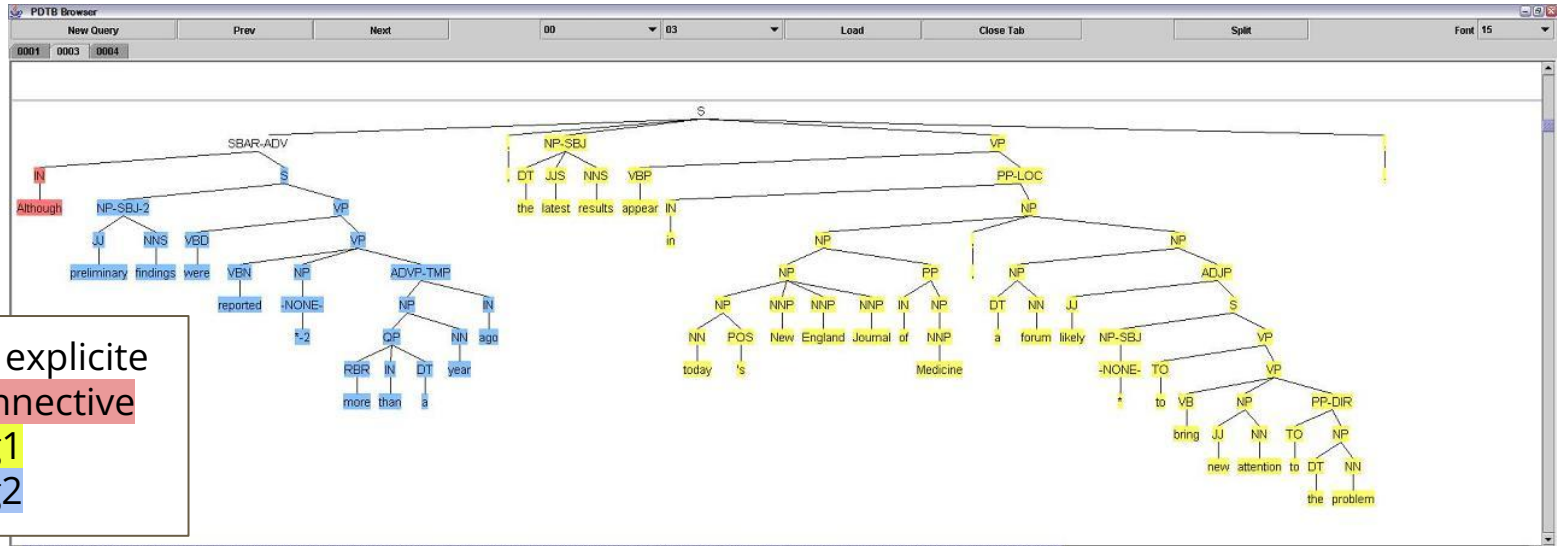
Cohérence locale : le Penn Discourse TreeBank

PDTB 2.0 (anglais) [Prasad et al., 2008]

- Basé sur les **connecteurs** et l'**adjacence** (relaxé dans le PDTB 3)
 - Explicite : 18,459 / Implicite : 16,224
 - Aussi Alternative lexicalizations, Entity relations et No relations: **40,600** au total
- Relations multiples (2% implicites)
- Distinction Arg1 / Arg2 (contenant le connecteur)
- **Plat / pas de structure**
- Hiérarchie de relations : **4 classes, 16 types, 23 sous-types**
- **2,259 WSJ articles** (<PTB)

| Mrs Yeargin is lying. | 1 **Implicit = because** | They found students in an advanced class a year earlier who said she gave them similar help. | 2 (CONTINGENCY:Pragmatic Cause:justification)

| The Bank of England, on the other hand, had gold reserves that averaged about 30% of its outstanding currency (...)| **AltLex = The most likely reason for this disparity is that** | (...)|



- Exemple explicite
- Connective
 - Arg1
 - Arg2

Conn: once

EntRel

Conn: Although

connHead	sClassA	Source	Type	Polarity	Det	rawText
although	Comparison Contrast	Wr	Comm	Null	Null	Although

Arg1: the latest results appear in L...

Arg2: preliminary findings were repo...

START

A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago, researchers reported.

The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said. Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1958.

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

A Lorillard spokeswoman said, "This is an old story. We're talking about years ago before anyone heard of asbestos having any questionable properties. There is no asbestos in our products now."

Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes. "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute. Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University.

The Lorillard spokeswoman said asbestos was used in "very modest amounts" in making paper for the filters in the early 1950s and replaced with a different type of filter in 1958. From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold, the company said.

Among 33 men who worked closely with the substance, 28 have died -- more than three times the expected number. Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer.

Cohérence locale : le Penn Discourse TreeBank

PDTB 2.0 (anglais) [Prasad et al., 2008]

- Basé sur les **connecteurs** et l'**adjacence** (relaxé dans le PDTB 3)
 - Explicite : 18,459 / Implicite : 16,224
 - Aussi Alternative lexicalizations, Entity relations et No relations: **40,600** au total → **NOT USED**
- Relations multiples (2% implicites) → **NOT USED**
- Distinction Arg1 / Arg2 (contenant le connecteur)
- **Plat / pas de structure**
- Hiérarchie de relations : **4 classes, 16 types, 23 sous-types** → **NOT USED** (très peu de résultats au niv. 3)
- **2,259 WSJ articles** (<PTB)

→ Très utilisé car gros, mais plutôt pour les sous-tâches mais :

- **Seulement partiellement exploité / simplifications**
- Pas clair si ces annotations sont utiles à des applications
- Plein de projets: Arabic, Czech, Chinese, Turkish, French (partial) (Hindi? Italian?) + TED Talks multilingual (2019)

Corpus	Domain	#Docs	#Sents	#Tokens	Vocab	#EDUs	#Conn	#Labels	#Rels	References
Tasks 1 and 3: EDU Segmentation and Relation Classification										
deu.rst.pcc	newspaper commentaries	176	2,193	33,222	8,359	3,018	-	26	2,665	Potsdam Commentary Corpus (Stede and Neumann, 2014)
**eng.dep.covdtb	scholarly paper abstracts on COVID-19 and related coronaviruses	300	2,343	60,849	8,293	5,705	-	12	4,985	COVID-19 Discourse Dependency Treebank (COVID19-DTB) (Nishida and Matsumoto, 2022)
eng.dep.scidtb	scientific articles	798	4,202	102,493	8,700	10,986	-	24	9,904	Discourse Dependency Treebank for Scientific Abstracts (SciDTB) (Yang and Li, 2018)
**eng.rst.gentle	multi-genre	26	1,334	17,797	4,135	2,708	-	31	2,540	Genre Tests for Linguistic Evaluation (GENTLE) (Aoyama et al., 2023)
eng.rst.gum	multi-genre	213	11,656	203,879	19,404	26,252	-	14	24,688	Georgetown University Multi-layer corpus V9 (Zeldes, 2017)
eng.rst.rstdt	news	385	8,318	205,829	19,160	21,789	-	17	19,778	RST Discourse Treebank (Carlson et al., 2001)
eng.sdrst.stac	dialogues	45	11,087	52,354	3,967	12,588	-	16	12,235	Strategic Conversations corpus (Asher et al., 2016)
eus.rst.ert	medical, terminological and scientific	164	2,380	45,780	13,662	4,202	-	29	3,825	Basque RST Treebank (Iruskieta et al., 2013)
fas.rst.prstc	journalistic texts	150	2,179	66,694	7,880	5,853	-	17	5,191	Persian RST Corpus (Shahmohammadi et al., 2021)
fra.sdrst.annodis	news, wiki	86	1,507	32,699	7,513	3,429	-	18	3,338	ANNOtation DIScursive (Afanteros et al., 2012).
nld.rst.nldt	expository texts and persuasive genres	80	1,651	24,898	4,935	2,343	-	32	2,264	Dutch Discourse Treebank (Redeker et al., 2012)
por.rst.cstn	news	140	2,221	58,793	7,786	5,537	-	32	4,993	Cross-document Structure Theory News Corpus (Cardoso et al., 2011)
rus.rst.rrt	blog and news	332	23,044	473,005	75,285	41,532	-	22	34,566	Russian RST Treebank (Toldova et al., 2017)
spa.rst.rststb	multi-genre	267	2,089	58,717	9,444	3,351	-	28	3,049	RST Spanish Treebank (da Cunha et al., 2011)
spa.rst.sctb	multi-genre	50	516	16,515	3,735	744	-	25	692	RST Spanish-Chinese Treebank (Spanish) (Cao et al., 2018)
zho.dep.scidtb	scientific	109	609	18,761	2,427	1,407	-	23	1,298	Discourse Dependency Treebank for Scientific Abstracts (SciDTB) (Yi et al., 2021; Cheng and Li, 2019)
zho.rst.gcdt	multi-genre	50	2,692	62,905	9,818	9,706	-	31	8,413	Georgetown Chinese Discourse Treebank (GCDT) (Peng et al., 2022b,a)
zho.rst.sctb	multi-genre	50	580	15,496	2,973	744	-	26	692	RST Spanish-Chinese Treebank (Chinese) (Cao et al., 2018)

RST and SDRT corpora

Tasks 2 and 3: Connective Detection and Relation Classification

eng.pdtb.gum	multi-genre	213	11,656	203,879	19,404	-	6,515	-	-	Georgetown University Multi-layer corpus V9 (Zeldes, 2017)
eng.pdtb.pdtb	news	2,162	48,630	1,156,657	48,937	-	26,048	23	47,851	Penn Discourse Treebank (Prasad et al., 2014; Webber et al., 2019b)
**eng.pdtb.tedm	TED talks	6	381	8,048	1,881	-	341	20	529	TED-Multilingual Discourse Bank (English) (Zeyrek et al., 2018, 2019)
ita.pdtb.luna	speech	60	3,753	26,114	2,392	-	1,071	15	1,544	LUNA Corpus Discourse Data Set (Tonelli et al., 2010; Riccardi et al., 2016)
por.pdtb.crpc ⁹	news, fiction, and didactic/scientific texts	302	5,194	186,849	22,208	-	5,159	19	11,330	Portuguese Discourse Bank (CRPC) (Mendes and Lejeune, 2022; Génèreux et al., 2012)
**por.pdtb.tedm	TED talks	6	394	8,190	2,162	-	305	20	554	TED-Multilingual Discourse Bank (Portuguese) (Zeyrek et al., 2018, 2019)
tha.pdtb.tdtb	news	180	6,534	256,523	11,789	-	10,864	21	10,865	Thai Discourse Treebank (TDTB)
tur.pdtb.tdb	multi-genre	197	31,196	487,389	88,923	-	8,748	23	3,185	Turkish Discourse Bank (Zeyrek and Webber, 2008; Zeyrek and Kurfali, 2017)
**tur.pdtb.tedm	TED talks	6	410	6,143	2,771	-	382	23	577	TED-Multilingual Discourse Bank (Turkish) (Zeyrek et al., 2018, 2019)
zho.pdtb.cdtb	news	164	2,891	73,314	9,085	-	1,660	9	5,270	Chinese Discourse Treebank (Zhou et al., 2014)

PDTB
corpora

Table 1: General Statistics of DISRPT Datasets: ** indicates an OOD dataset. ‘#Docs’, ‘#Sents’, ‘#Tokens’ and ‘#EDUs’ correspond resp. to the total number of documents, sentences (Treebanked track), tokens, and EDUs. #Conn is the number of tokens starting a connective, and ‘Vocab’ of unique tokens. ‘#Labels’ is the size of the respective label set and ‘#Rels’ to the total number of pairs annotated.

L'analyse discursive automatique

Beaucoup d'applications de TAL se limitent au niveau de la phrase :

- mais problématique, cf exemple du résumé automatique
- le contexte est aussi crucial en traduction automatique, extraction d'info, chatbots...

L'information discursive est nécessaire (si nécessaire) à la fin de la chaîne :

- signifie aussi : besoin des informations des étapes précédentes (lesquelles ? pas clair)
- et qu'il faut gérer la propagation d'erreurs

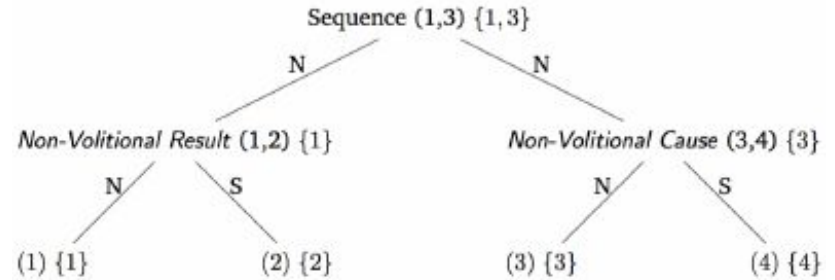
Les systèmes dédiés à l'analyse discursive s'intéressent à :

- **Full discourse parsing**, ce qui ne signifie pas la même chose en RST ou sur le PDTB
- Les sous-tâches, e.g. identification des connecteurs, segmentation, identification des relations, attachement ...

Discourse parsing: 2 views

Full discourse parsing (RST, SDRT)

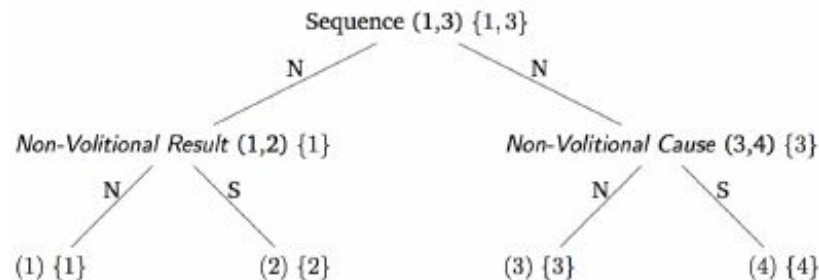
[The towers collapsed less than two hours later]1 (Result)
[dragging down with them the building of the Marriott
World Trade Center.]2 (Sequence) [The tower 7 of the
WTC collapsed in the afternoon]3 [because of damages
caused by the fall of Twin Towers.]4



Discourse parsing: 2 views

Full discourse parsing (RST, SDRT)

[The towers collapsed less than two hours later]1 (Result)
[dragging down with them the building of the Marriott
World Trade Center.]2 (Sequence) [The tower 7 of the
WTC collapsed in the afternoon]3 [because of damages
caused by the fall of Twin Towers.]4



Shallow discourse parsing / Discourse chunking (PDTB)

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

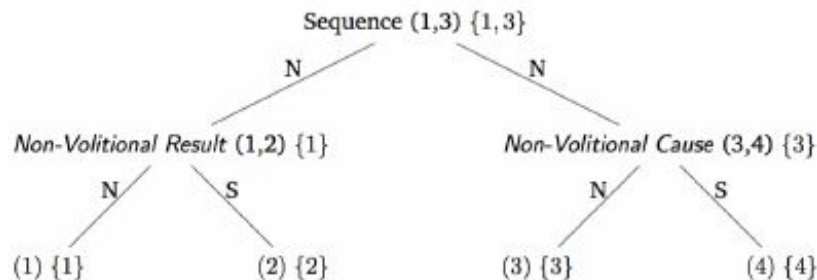
A Lorillard spokeswoman said, "This is an old story, We're talking about years ago before anyone heard of asbestos having any questionable properties. There is no asbestos in our products now."

Conn: Although	
connHead	sClassA
although	Comparison.Contrast

Discourse parsing: 2 views

Full discourse parsing (RST, SDRT)

[The towers collapsed less than two hours later]1 (Result)
[dragging down with them the building of the Marriott
World Trade Center.]2 (Sequence) [The tower 7 of the
WTC collapsed in the afternoon]3 [because of damages
caused by the fall of Twin Towers.]4



Shallow discourse parsing / Discourse chunking (PDTB)

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

A Lorillard spokeswoman said, "This is an old story, We're talking about years ago before anyone heard of asbestos having any questionable properties.

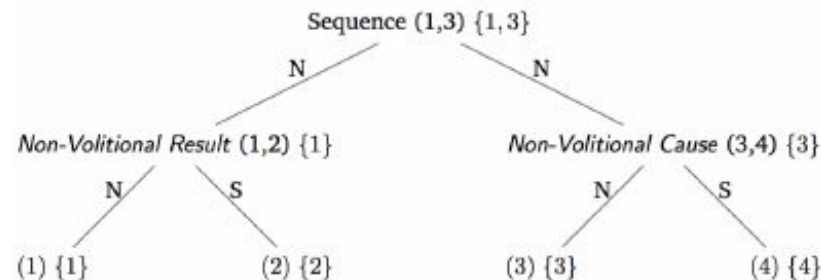
There is no asbestos in our products now."

Implicit: in fact	
conn1	sClass1A
in fact	Expansion.Restatement.Specification

Discourse parsing: 2 views

Full discourse parsing (RST, SDRT)

[The towers collapsed less than two hours later]1 (Result)
[dragging down with them the building of the Marriott
World Trade Center.]2 (Sequence) [The tower 7 of the
WTC collapsed in the afternoon]3 [because of damages
caused by the fall of Twin Towers.]4



Shallow discourse parsing / Discourse chunking (PDTB)

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

A Lorillard spokeswoman said, "This is an old story,

We're talking about years ago before anyone heard of asbestos having any questionable properties.

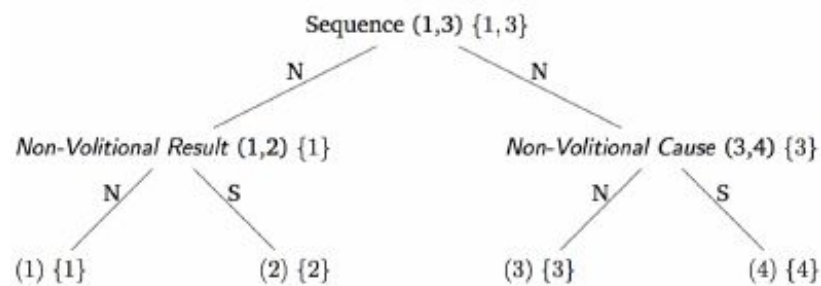
There is no asbestos in our products now."

Implicit: besides		
conn1	sClass 1A	sClass 1B
besides	Expansion.Conjunction	Comparison

Parsing discursif (RST ou SDRT)

Etapes :

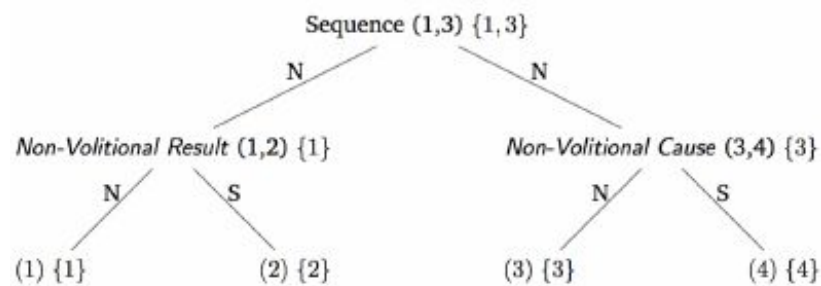
1. Segmentation en Elementary Discourse Units (EDU)
 2. Tree Building = processus récursif : construction de la structure à partir des unités de discours (uniquement EDU au début)
 - a. attachement : attacher une paire d'unités
 - b. identification de la relation : étiqueter la relation entre les unités attachées
 - c. + bonus RST : étiquetage des segments entre nucleus ou satellite
 - d. recommencer : la paire d'unités liées devient une nouvelle unité (complexe)
- on s'arrête quand on a couvert tous les segments de texte



Parsing discursif (RST ou SDRT)

Etapas :

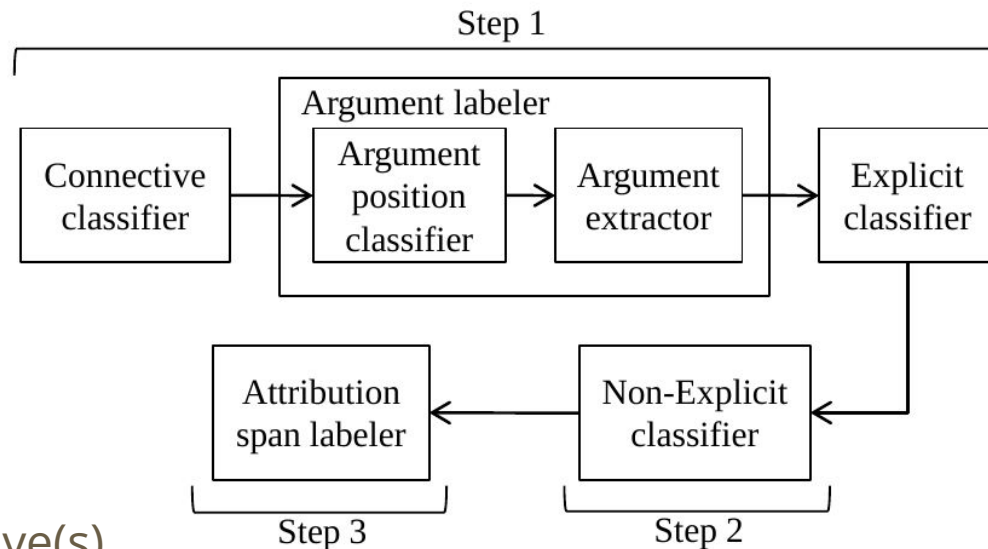
1. **Segmentation** en Elementary Discourse Units (EDU)
2. Tree building = processus récursif : construction de la structure à partir des unités de discours (uniquement EDU au début)
 - a. **attachement** : attacher une paire d'unités
 - b. **identification de la relation** : étiqueter la relation entre les unités attachées
 - c. + bonus RST : étiquetage des segments entre nucleus ou satellite
 - d. recommencer : la paire d'unités liées devient une nouvelle unité (complexe)→ on s'arrête quand on a couvert tous les segments de texte



Chunking discursif ou shallow parsing (PDTB)

Pipeline :

1. Identifier les connecteurs
2. Identifier leurs arguments
 - a. position (inter / intra, ordre)
 - b. frontières
3. Identifier la / les relation discursive(s)
 - a. explicites
 - b. implicites
 - c. (entRel, AltLex, NoRel)



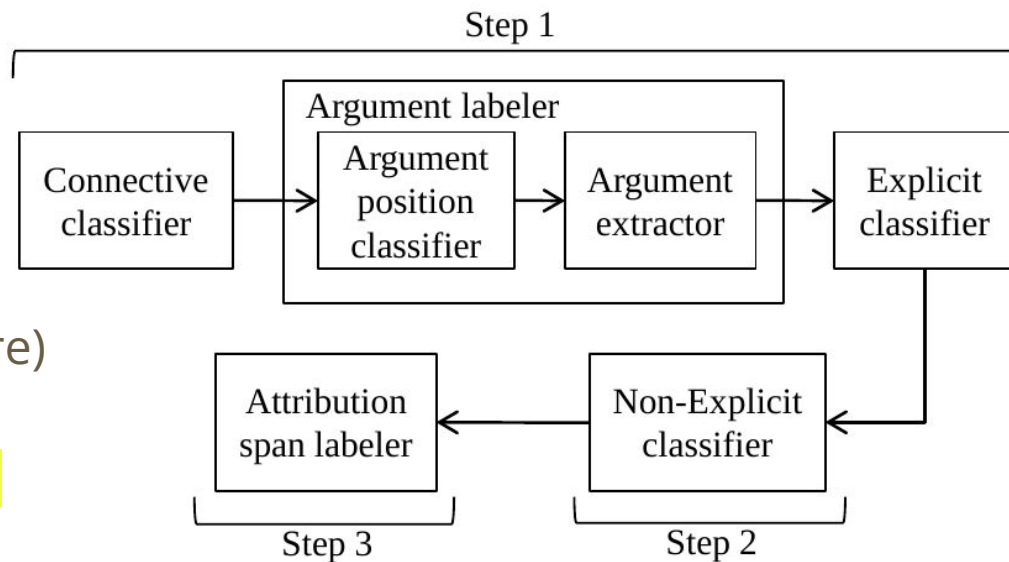
Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

Implicit: besides		
conn1	sClass1A	sClass1B
besides	Expansion.Conjunction	Comparison

Chunking discursif ou shallow parsing (PDTB)

Pipeline :

1. Identifier les connecteurs
2. Identifier leurs arguments
 - a. position (inter / intra, ordre)
 - b. frontières
3. Identifier la relation discursive
 - a. explicites
 - b. implicites
 - c. (entRel, AltLex, NoRel)



Différentes applications

- Co-référence [Cristea et al., 1999]
- Automatic summarization [Sporleder and Lapata, 2005; Gerani et al., 2014; Schrimpf, 2018, Koto et al, 2019]
- Question Answering [Verberne, 2007]
- Sentiment analysis [Bhatia et al., 2015; Lee et al. 2018]
- Essay scoring [Higgins et al., 2004; Mesgar and Strube, 2018]
- Summary coherence rating [Nguyen and Joty, 2017], coherence Modeling [Li and Jurafsky, 2017; Mesgar and Strube, 2018]
- Machine translation [Meyer and Webber, 2013; Born et al., 2017]
- Popularity prediction [Koto et al, 2019]
- Sarcasm detection [Lee et al. 2018]
- Alzheimer analysis [Abdalla et al. 2018]
- Fraud detection [Braud and Soegaard 2017]
- Schizophrenia and depression analysis [Li et al 2020, Li et al 2022]

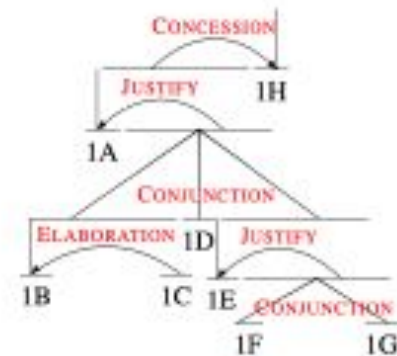
Sentiment analysis [Bahtia et al. 2015]

Méthodes:

1. Repondération de la contribution de chaque unité de discours basée sur sa position, combiné à :

- un analyseur de sentiment basé sur un lexique
- ou un classifieur

2. Propagation récursive du sentiment de bas en haut de l'arbre discours RST (RecNN)



[It could have been a great movie]^{1A} [It does have beautiful scenery,]^{1B} [some of the best since Lord of the Rings.]^{1C} [The acting is well done,]^{1D} [and I really liked the son of the leader of the Samurai.]^{1E} [He was a likable chap,]^{1F} [and I hated to see him die.]^{1G} [But, other than all that, this movie is nothing more than hidden rip-offs.]^{1H}

	Pang & Lee	Socher et al.
<i>Baselines</i>		
B1. Lexicon	68.3	74.9
B2. Classifier	82.4	81.5
<i>Discourse depth weighting</i>		
D1. Lexicon	72.6	78.9
D2. Classifier	82.9	82.0
<i>Rhetorical recursive neural network</i>		
R1. No relations	83.4	85.5
R2. With relations	84.1	85.6

Summarization DiscoBERT [Xu et al. 2020]

- Segmentation en EDUs plutôt que phrases : plus fin, moins redondant
- Graphes RST : dépendances à longue portée entre EDUs [Ji and Eisenstein, 2014]
- Graphes de coréférence : modélisation d'un contexte large
- Encoder de document basé sur BERT, Graph Convolutional Networks pour encoder les graphes

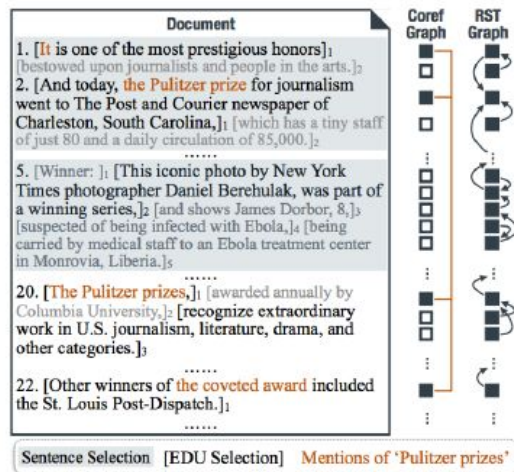


Figure 1: Illustration of DISCOBERT for text summarization. Sentence-based BERT model (baseline) selects whole sentences 1, 2 and 5. The proposed discourse-aware model DISCOBERT selects EDUs {1-1, 2-1, 5-2, 20-1, 20-3, 22-1}. The right side of the figure illustrates the two discourse graphs we use: (i) Coref(erence) Graph (with the mentions of 'Pulitzer prizes' highlighted as examples); and (ii) RST Graph (induced by RST discourse trees).

Model	R-1	R-2	R-L
Lead3	40.42	17.62	36.67
Oracle (Sentence)	55.61	32.84	51.88
Oracle (Discourse)	61.61	37.82	59.27
NeuSum (Zhou et al., 2018)	41.59	19.01	37.98
BanditSum (Dong et al., 2018)	41.50	18.70	37.60
JECS (Xu and Durrett, 2019)	41.70	18.50	37.90
PNBERT (Zhong et al., 2019)	42.39	19.51	38.69
PNBERT w. RL	42.69	19.60	38.85
BERT (Zhang et al., 2019)	41.82	19.48	38.30
HIBERT _S	42.10	19.70	38.53
HIBERT _S *	42.31	19.87	38.78
HIBERT _M *	42.37	19.95	38.83
BERTSUM (Liu and Lapata, 2019)	43.25	20.24	39.63
T5-Base (Raffel et al., 2019)	42.05	20.34	39.40
BERT	43.07	19.94	39.44
DISCOBERT	43.38	20.44	40.21
DISCOBERT w. \mathcal{G}_C	43.58	20.64	40.42
DISCOBERT w. \mathcal{G}_R	43.68	20.71	40.54
DISCOBERT w. \mathcal{G}_R & \mathcal{G}_C	43.77	20.85	40.67

Table 2: Results on the test set of the CNNDM dataset. ROUGE-1, -2 and -L F_1 are reported. Models with the asterisk symbol (*) used extra data for pre-training. R-1 and R-2 are shorthands for unigram and bigram overlap; R-L is the longest common subsequence.



AnDiAMO

2022 - 2025

Analysing **D**iscourse **A**utomatically, with **M**ultiple **O**bjectives

- Descriptif du projet
- Questions scientifiques
- Méthodologie
- Collaborateurs

AnDiAMO

- ANR Jeunes Chercheur(e)s
- 36 mois : mars 2022 - mars 2025

Recrutement :

- postdoc : 18 mois (Eleni Metheniti)
- ingénieure de recherche : 18 mois (Laura Rivière)
- stages de master : 18 mois

Projet de Recherche

Goal = Improve performance of discourse parsers

→ Challenges:

- data scarcity, even for English, worse for other languages
- complex problem: even modern large pretrained language models are struggling

→ Approaches:

- Explore cross-lingual settings: increase data available + develop models for low resource languages
- Explore cross-framework approaches: increase data available + allow to understand differences between frameworks / annotation projects
- ++ Evaluation: how to use discourse information for downstream tasks?

Projet de Recherche

Méthodologie :

- More on the machine learning side:
 - Representation learning: based on Pretrained Large Models, incl. feature augmentation
 - Multi-task learning
- More on the linguistic side:
 - work on relation sets
 - work on combining languages / corpora

RST

- **Attribution:** attribution, attribution-negative
- **Background:** background, circumstance
- **Cause:** cause, result, consequence
- **Comparison:** comparison, preference, analogy, proportion
- **Condition:** condition, hypothetical, contingency, otherwise
- **Contrast:** contrast, concession, antithesis
- **Elaboration:** elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
- **Enablement:** purpose, enablement
- **Evaluation:** evaluation, interpretation, conclusion, comment
- **Explanation:** evidence, explanation-argumentative, reason
- **Joint:** list, disjunction
- **Manner-Means:** manner, means
- **Topic-Comment:** problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
- **Summary:** summary, restatement
- **Temporal:** temporal-before, temporal-after, temporal-same-time, sequence, inverted sequence
- **Topic Change**:** topic-shift, topic-drift

PDTB

• TEMPORAL

- Asynchronous
- Synchronous:
precedence, succession

• CONTINGENCY

- Cause: result, reason
- Pragmatic cause:
justification
- Condition: hypothetical, general, unreal present, unreal past, real present, real past
- Pragmatic condition:
relevance, implicit assertion

• COMPARISON

- Contrast: juxtaposition, opposition
- Pragmatic contrast
- Concession: expectation, contra-expectation
- Pragmatic concession

• EXPANSION

- Conjunction
- Instantiation
- Restatement: specification, equivalence, generalization
- Alternative: conjunctive, disjunctive, chosen alternative
- Exception
- List

- **Attribution:** attribution, attribution-negative
- **Background:** background, circumstance
- **Cause:** cause, result, consequence
- **Comparison:** comparison, preference, analogy, proportion
- **Condition:** condition, hypothetical, contingency, otherwise
- **Contrast:** contrast, concession, antithesis
- **Elaboration:** elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
- **Enablement:** purpose, enablement
- **Evaluation:** evaluation, interpretation, conclusion, comment
- **Explanation:** evidence, explanation-argumentative, reason
- **Joint:** list, disjunction
- **Manner-Means:** manner, means
- **Topic-Comment:** problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
- **Summary:** summary, restatement
- **Temporal:** temporal-before, temporal-after, temporal-same-time, sequence, inverted sequence
- **Topic Change**:** topic-shift, topic-drift

PDTB

• TEMPORAL

- Asynchronous
- Synchronous: precedence, succession

• CONTINGENCY

- Cause: result, reason
- Pragmatic cause: justification
- Condition: hypothetical, general, unreal present, unreal past, real present, real past
- Pragmatic condition: relevance, implicit assertion

• COMPARISON

- Contrast: juxtaposition, opposition
- Pragmatic contrast
- Concession: expectation, contra-expectation
- Pragmatic concession

• EXPANSION

- Conjunction
- Instantiation
- Restatement: specification, equivalence, generalization
- Alternative: conjunctive, disjunctive, chosen alternative
- Exception
- List

Travaux réalisés / en cours

- Données : le benchmark DISRPT
- Multilingue: segmentation, connecteurs
- Also cross-framework: relations
- Full Parsing
- Distant learning: downstream applications

DISRPT: a benchmark for discourse processing

Historique :

- 2007-2017 : Workshops centrés sur la RST
- 2019 : [1st DISRPT](#)
 - élargir la portée du workshop pour inclure les autres frameworks
 - 1ère shared task: segmentation + connecteurs en RST, SDRT, PDTB pour 15 corpus
- 2021 : [2nd shared task DISRPT](#) (workshop CODI), [\[Zeldes et al 2021\]](#)
 - segmentation + connecteurs + **relations** pour 16 jeux de données
- 2023 : [3rd shared task DISRPT](#), [\[Braud et al 2023\]](#)
 - segmentation + connecteurs + relations pour **26 jeux de données**
 - → Release 'officielle' du jeu de données DISRPT sur HuggingFace

Why an unified benchmark?

L'annotation en discours, c'est un peu le Far West ... Chaque nouveau projet met en place de nouvelles règles (et une dizaine de formats de fichiers différents !!) :

- limite les études cross-lingues car on a rarement 2 corpus avec les mêmes règles d'annotation pour différentes langues
- limite aussi du coup l'évaluation de la robustesse des systèmes automatique à travers les langues ou domaines
- augmente (artificiellement) le problème du manque de données car on ne peut pas facilement combiner les corpus... pourtant on en a 'beaucoup', environ 4 millions de tokens annotés

Idéalement on voudrait un format **unifié sur tous les frameworks, facile à utiliser** et **fidèle à l'annotation originale**

DISRPT: a unified benchmark [LREC under review]

- Problème de framework divergents :
 - pas de segmentation dans le PDTB / pas de connecteurs dans les RST, SDRT(, DEP) → **tâches séparés selon les corpus**
 - segmentation enchâssée dans la SDRT mais pas RST (, DEP) → cf exemple, **choix du modèle RST**
 - jeux de relations très différents → un peu d'**homogénéisation, pas d'unification**
- Deux formats :
 - CoNLL-U : segmentation / connecteur = classification au niveau du token
 - RELS : label entre des paires de segments

ex **RST**: [But maintaining the key components (. . .)]1 [- a stable exchange rate and high levels of imports -]2 [will consume enormous amounts (. . .).]3 (Carlson and Marcu, 2001)

SDRT: [But maintaining the key components (. . .) [- a stable exchange rate and high levels of imports -]2 will consume enormous amounts (. . .).]1

```

# newdoc_id = wsj_0605
1 UNDER under ADP IN _ 3 case 3:case Seg=B-seg
2 A a DET DT Definite=Ind|PronType=Art 3 det 3:det Seg=0
3 PROPOSAL proposal NOUN NN Number=Sing 25 obl 25:obl:under Seg=0
4 by by ADP IN _ 5 case 5:case Seg=0
5 Democrats Democrats PROPN NNPS Number=Plur 3 nmod 3:nmod:by Seg=0
6 to to PART TO _ 7 mark 7:mark Seg=B-seg
7 expand expand VERB VB VerbForm=Inf 3 acl 3:acl:to Seg=0
8 Individual Individual PROPN NNP Number=Sing 10 compound 10:compound Seg=0
9 Retirement Retirement PROPN NNP Number=Sing 10 compound 10:compound Seg=0
10 Accounts Accounts PROPN NNPS Number=Plur 7 obj 7:obj Seg=0
11 , , PUNCT , _ 25 punct 25:punct Seg=0
12 a a DET DT Definite=Ind|PronType=Art 15 det 15:det Seg=B-seg
13 $ $ SYM $ _ 15 amod 15:amod Seg=0
14 2,000 2,000 NUM CD NumForm=Digit|NumType=Card 13 compound 13:compound Seg=0
15 contribution contribution NOUN NN Number=Sing 25 nsubj 25:nsubj Seg=0
16 by by ADP IN _ 18 case 18:case Seg=0
17 a a DET DT Definite=Ind|PronType=Art 18 det 18:det Seg=0
18 taxpayer taxpayer NOUN NN Number=Sing 15 nmod 15:nmod:by Seg=0

```

Segmentation

```

20 as as ADV RB _ 21 advmod 21:advmod Conn=B-conn
21 soon soon ADV RB Degree=Pos 16 advmod 16:advmod Conn=I-conn
22 as as SCONJ IN _ 27 mark 27:mark Conn=I-conn
23 the the DET DT Definite=Def|PronType=Art 26 det 26:det Conn=0
24 New New PROPN NNP Number=Sing 26 compound 26:compound Conn=0
25 York York PROPN NNP Number=Sing 26 compound 26:compound Conn=0
26 market market NOUN NN Number=Sing 27 nsubj 27:nsubj Conn=0

```

Connecteur

Corpus	Domain	#Docs	#Sents	#Tokens	Vocab	#EDUs	#Conn	#Labels	#Rels	References
Tasks 1 and 3: EDU Segmentation and Relation Classification										
deu.rst.pcc	newspaper commentaries	176	2,193	33,222	8,359	3,018	-	26	2,665	Potsdam Commentary Corpus (Stede and Neumann, 2014)
**eng.dep.covdtb	scholarly paper abstracts on COVID-19 and related coronaviruses	300	2,343	60,849	8,293	5,705	-	12	4,985	COVID-19 Discourse Dependency Treebank (COVID19-DTB) (Nishida and Matsumoto, 2022)
eng.dep.scidtb	scientific articles	798	4,202	102,493	8,700	10,986	-	24	9,904	Discourse Dependency Treebank for Scientific Abstracts (SciDTB) (Yang and Li, 2018)
**eng.rst.gentle	multi-genre	26	1,334	17,797	4,135	2,708	-	31	2,540	Genre Tests for Linguistic Evaluation (GENTLE) (Aoyama et al., 2023)
eng.rst.gum	multi-genre	213	11,656	203,879	19,404	26,252	-	14	24,688	Georgetown University Multi-layer corpus V9 (Zeldes, 2017)
eng.rst.rstdt	news	385	8,318	205,829	19,160	21,789	-	17	19,778	RST Discourse Treebank (Carlson et al., 2001)
eng.sdrst.stac	dialogues	45	11,087	52,354	3,967	12,588	-	16	12,235	Strategic Conversations corpus (Asher et al., 2016)
eus.rst.ert	medical, terminological and scientific	164	2,380	45,780	13,662	4,202	-	29	3,825	Basque RST Treebank (Iruskieta et al., 2013)
fas.rst.prstc	journalistic texts	150	2,179	66,694	7,880	5,853	-	17	5,191	Persian RST Corpus (Shahmohammadi et al., 2021)
fra.sdrst.annodis	news, wiki	86	1,507	32,699	7,513	3,429	-	18	3,338	ANNOtation DIScursive (Afanteros et al., 2012).
nld.rst.nldt	expository texts and persuasive genres	80	1,651	24,898	4,935	2,343	-	32	2,264	Dutch Discourse Treebank (Redeker et al., 2012)
por.rst.cstn	news	140	2,221	58,793	7,786	5,537	-	32	4,993	Cross-document Structure Theory News Corpus (Cardoso et al., 2011)
rus.rst.rrt	blog and news	332	23,044	473,005	75,285	41,532	-	22	34,566	Russian RST Treebank (Toldova et al., 2017)
spa.rst.rststb	multi-genre	267	2,089	58,717	9,444	3,351	-	28	3,049	RST Spanish Treebank (da Cunha et al., 2011)
spa.rst.sctb	multi-genre	50	516	16,515	3,735	744	-	25	692	RST Spanish-Chinese Treebank (Spanish) (Cao et al., 2018)
zho.dep.scidtb	scientific	109	609	18,761	2,427	1,407	-	23	1,298	Discourse Dependency Treebank for Scientific Abstracts (SciDTB) (Yi et al., 2021; Cheng and Li, 2019)
zho.rst.gcdt	multi-genre	50	2,692	62,905	9,818	9,706	-	31	8,413	Georgetown Chinese Discourse Treebank (GCDT) (Peng et al., 2022b,a)
zho.rst.sctb	multi-genre	50	580	15,496	2,973	744	-	26	692	RST Spanish-Chinese Treebank (Chinese) (Cao et al., 2018)

RST and SDRT corpora

Tasks 2 and 3: Connective Detection and Relation Classification

eng.pdtb.gum	multi-genre	213	11,656	203,879	19,404	-	6,515	-	-	Georgetown University Multi-layer corpus V9 (Zeldes, 2017)
eng.pdtb.pdtb	news	2,162	48,630	1,156,657	48,937	-	26,048	23	47,851	Penn Discourse Treebank (Prasad et al., 2014; Webber et al., 2019b)
**eng.pdtb.tedm	TED talks	6	381	8,048	1,881	-	341	20	529	TED-Multilingual Discourse Bank (English) (Zeyrek et al., 2018, 2019)
ita.pdtb.luna	speech	60	3,753	26,114	2,392	-	1,071	15	1,544	LUNA Corpus Discourse Data Set (Tonelli et al., 2010; Riccardi et al., 2016)
por.pdtb.crpc ⁹	news, fiction, and didactic/scientific texts	302	5,194	186,849	22,208	-	5,159	19	11,330	Portuguese Discourse Bank (CRPC) (Mendes and Lejeune, 2022; Généréux et al., 2012)
**por.pdtb.tedm	TED talks	6	394	8,190	2,162	-	305	20	554	TED-Multilingual Discourse Bank (Portuguese) (Zeyrek et al., 2018, 2019)
tha.pdtb.tdtb	news	180	6,534	256,523	11,789	-	10,864	21	10,865	Thai Discourse Treebank (TDTB)
tur.pdtb.tdb	multi-genre	197	31,196	487,389	88,923	-	8,748	23	3,185	Turkish Discourse Bank (Zeyrek and Webber, 2008; Zeyrek and Kurfali, 2017)
**tur.pdtb.tedm	TED talks	6	410	6,143	2,771	-	382	23	577	TED-Multilingual Discourse Bank (Turkish) (Zeyrek et al., 2018, 2019)
zho.pdtb.cdtb	news	164	2,891	73,314	9,085	-	1,660	9	5,270	Chinese Discourse Treebank (Zhou et al., 2014)

PDTB
corpora

Table 1: General Statistics of DISRPT Datasets: ** indicates an OOD dataset. ‘#Docs’, ‘#Sents’, ‘#Tokens’ and ‘#EDUs’ correspond resp. to the total number of documents, sentences (Treebanked track), tokens, and EDUs. #Conn is the number of tokens starting a connective, and ‘Vocab’ of unique tokens. ‘#Labels’ is the size of the respective label set and ‘#Rels’ to the total number of pairs annotated.

Segmentation discursive

- Considérée comme une tâche facile et résolue 93.7% [Xuan Bach et al, 2012] mais :
 - quasi uniquement pour l'anglais, les monologues et le WSJ, pas de comparaisons entre langues, modalités, cadres théoriques...
 - nécessite une segmentation gold en phrase (pas si facile que ça ...)
 - nécessite une analyse syntaxique (semble cher pour une simple 'tokenisation')
- D'où :
 - [Cross-lingual and cross-domains discourse segmentation](#). (2017)
 - [Does syntax helps discourse segmentation? Not so much](#). (2017)
 - [ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents](#) (2019)
 - [Multi-lingual Discourse Segmentation and Connective Identification: MELODI at Disrpt2021](#) (2021)
 - [DisCut and DiscReT: MELODI at DISRPT 2023](#) (2023)

DISRPT
shared
task

Segmentation discursive = segmenter un document en EDUs

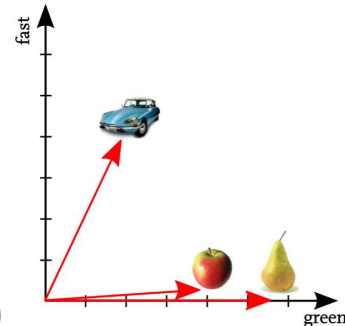
- Clauses, max phrase, mais RST DT = grain fin cf les règles [tagging manual](#) :
 - Includes both speech acts and other cognitive acts:
 - (a) |The company **says** | |it will shut down its plant.|
 - (b) |HDTV is where everybody is going,| |**says** ...|
 - But if the complement is a to-infinitival, do not segment:
 - (c) |The company wants **to** shut down its plant.|
 - But segment infinitive clause marking a purpose relation:
 - (d) |A grand jury has been investigating whether officials (...) conspired **to** cover up their accounting | |**to** evade federal income taxes. |
- Autres cas :
 - (e) |Mr. Volk, 55 years old, succeeds Duncan Dwight, | |**who** retired in September. | (unité enchâssée, on ne modifie qu'une partie de l'EDU, mais considérée de la même manière)
 - (f) |The Tass news agency said the 1990 budget anticipates income of 429.9 billion rubles | | (**\$US693.4 billion**) |
- Segments enchâssés, vus comme 3 segments
 - (g) [But maintaining the key components (. . .)]_1 [**a stable exchange rate and high levels of imports** -]_2 [will consume enormous amounts (. . .)]_3

Word embeddings

Idée: les ordinateurs ne comprennent que les nombres → représenter les mots par des valeurs numériques, plus précisément des vecteurs $\langle f_1, f_2, f_3, \dots, f_n \rangle$

Sémantique distributionnelle : on connaît un mot aux mots qui l'entourent = on 'calcule le sens d'un mot' = vecteur en regardant avec quels mots il apparaît souvent

- a pu se faire par des calculs de fréquence en corpus
- aujourd'hui : **modèles neuronaux** entraînés de manière **non supervisé** pour apprendre ces vecteurs, e.g. deviner le mot suivant
 - 1ère époque : les embeddings statiques → 1 mot = 1 vecteur
 - Aujourd'hui : embeddings dynamiques, prennent en compte le context de la phrase = **Pretrained (Large) Language Models (PLM)**



→ Nécessite de grandes quantités de données, d'importantes ressources computationnelles

→ Modèles généraux : doivent être affiné sur une tâche particulière = fine-tuned

2019 : Comparaison sur les corpus anglais

	Rand 50d	GloVe 50d	BERT-E	BERT-M	ELMo
PDTB	77.08	65.17	90.83	89.89	88.40
GUM	80.58	78.28	86.29	87.27	87.65
RST DT	78.97	83.21	94.41	93.72	94.75
STAC	77.43	71.70	84.65	84.45	86.06

- Modèles dynamiques / contextuels >> modèle statiques = GloVe
- BERT multilingue quasi au niveau du monolingue sur l'anglais

Meilleur système en 2019 : <https://sites.google.com/view/disrpt2019/shared-task?authuser=0>

2021: grouper des langues similaires

→ résultats plus faibles sur des corpus plus petits: deux groupes de langue = romanes et germaniques

Méthode : Fusionner les corpus à l'entraînement, e.g.

- les 2 corpus espagnols
- les 2 corpus espagnols + le corpus portugais
- tous les corpus de langue romane (incl. français)
- + on test un fine-tuning supplémentaire pour le groupe espagnol / portugais sur les corpus cibles individuellement

Mais notre système est classé 3ème, cf [results here](#)

Corpus	Group	P	R	F1
spa/rststb	self	91.80	93.56	92.67
	SPA	94.24	93.79	94.02
	SPO	93.81	94.03	93.92
	ROM	93.66	91.65	92.64
	FT	94.30	94.75	94.52
spa/sctb	self	85.57	80.58	83.00
	SPA	83.64	89.32	86.38
	SPO	86.24	91.26	88.68
	ROM	81.98	88.35	85.05
	FT	88.89	85.44	87.13
por/cstn	self	90.63	92.06	91.34
	SPO	92.88	89.05	90.92
	ROM	90.14	90.00	90.07
	FT	90.98	92.86	91.91
fra/annodis	self	87.13	90.11	88.59
	ROM	91.12	83.09	86.92

Table 4: Intra sentential results on romance development datasets, with different training setups: SPA

Les stratégies de fusion permettent des améliorations par rapport à 'self' (baseline, pas de fusion)

Corpus	Group	P	R	F1
deu.rst.pcc	self	92.58	95.27	93.91
	GD	95.51	92.73	94.10
	GER	96.64	94.18	95.40
nld.rst.nldt	self	95.52	93.29	94.40
	GD	96.95	92.71	94.78
	GER	94.44	94.17	94.31
eng.rst.gum	self	94.06	90.19	92.08
	ENG	94.08	92.64	93.36
	GER	94.11	92.42	93.26
	FT	93.63	92.12	92.87
eng.rst.rstdt	self	96.35	95.38	95.86
	ENG	96.27	94.77	95.51
	GER	95.47	95.10	95.29
eng.sdrt.stac	FT	94.96	96.55	95.75
	self	94.19	95.49	94.84
	ENG	95.80	92.81	94.28
	GER	95.98	93.07	94.50
	FT	95.55	94.97	95.26

Table 5: Intra sentential results on germanic development datasets, with different training setups: GD

2023 : PLM multilingue segmentation + connecteurs

Tester simplement un gros modèle multilingue (architecture Transformer) :

- **BERTM-Base** : 104 languages, 12-layers, 768-hidden, 12-heads, **110M parameters**.
- **xlm-RoBERTa-base** : 100 languages, 12-layers, 768-hidden-state, 3072 feed-forward hidden-state, 8-heads, **125M parameters**
- **xlm-RoBERTa-large** : 24-layers, 1027-hidden-state, 4096 feed-forward hidden-state, 16-heads, **355M parameters**

Corpus	Treebanked			Plain		
	F1 dev	F1 test	DD21	F1 dev	F1 test	DD21
Segmentation						
deu.rst.pcc	96.79	96.01	95.58	96.60	94.24	93.94
eng.rst.gum	95.54	95.50	94.15	95.78	94.46	92.61
eng.rst.rstdt	97.33	97.62	96.64	97.60	97.74	96.35
eus.rst.ert	91.69	89.93	90.46	91.83	91.09	90.47
fas.rst.prstc	93.79	93.40	92.94	94.05	93.36	92.86
nld.rst.nldt	97.51	96.54	95.97	97.09	97.19	94.69
por.rst.cstn	94.06	93.98	94.35	93.50	94.36	94.11
rus.rst.rst	86.80	85.58	86.21	84.75	85.41	85.74
spa.rst.rststb	96.19	93.53	92.22	96.32	93.70	91.76
spa.rst.sctb	86.88	85.63	82.48	85.44	84.21	80.86
zho.rst.gcdt	92.69	92.55	-	92.20	91.74	-
zho.rst.sctb	79.05	81.84	83.34	77.53	78.55	76.21
eng.sdrst.stac	94.77	95.22	94.91	91.57	90.67	91.91
fra.sdrst.annodis	90.27	88.21	90.02	90.17	90.89	85.78
*eng.dep.covdtb	91.32	92.13	-	91.65	92.13	-
eng.dep.scidtb	96.18	95.07	-	95.63	94.49	-
zho.dep.scidtb	93.33	89.07	-	93.01	90.04	-
Mean	92.60	91.87	-	92.04	91.43	-
Mean corpora 2021	-	91.77	91.48	-	91.22	89.79
Connective identification						
eng.pdtb.pdtb	94.41	93.66	92.02	93.94	91.64	92.56
*eng.pdtb.tedm	75.86	78.36	-	80.00	75.83	-
ita.pdtb.luna	79.72	65.85	-	74.19	71.60	-
por.pdtb.crpc	85.16	80.66	-	84.65	79.49	-
*por.pdtb.tedm	73.08	80.29	-	71.22	79.45	-
*tha.pdtb.tdtb	87.43	85.66	-	74.32	69.92	-
tur.pdtb.tdb	89.73	92.77	94.11	89.69	91.12	93.56
*tur.pdtb.tedm	65.42	64.10	-	64.15	64.78	-
zho.pdtb.cdtb	87.66	89.00	87.52	87.77	90.38	88.35
Mean	82.05	81.15	-	79.99	79.36	-
Mean corpora 2021	-	91.81	91.22	-	91.05	91.49

2023 : PLM multilingue segmentation + connecteurs

Tester simplement un gros modèle multilingue (architecture Transformer) :

- **BERTM-Base** : 104 languages, 12-layers, 768-hidden, 12-heads, **110M parameters**. [Devlin et al 2019]
- **xlm-RoBERTa-base** : 100 languages, 12-layers, 768-hidden-state, 3072 feed-forward hidden-state, 8-heads, **125M parameters**
- **xlm-RoBERTa-large** : 24-layers, 1027-hidden-state, 4096 feed-forward hidden-state, 16-heads, **355M parameters**

Corpus	Treebanked			Plain		
	F1 dev	F1 test	DD21	F1 dev	F1 test	DD21
Segmentation						
deu.rst.pcc	96.79	96.01	95.58	96.60	94.24	93.94
eng.rst.gum	95.54	95.50	94.15	95.78	94.46	92.61
eng.rst.rstdt	97.33	97.62	96.64	97.60	97.74	96.35
eus.rst.ert	91.69	89.93	90.46	91.83	91.09	90.47
fas.rst.prstc	93.79	93.40	92.94	94.05	93.36	92.86
nld.rst.nldt	97.51	96.54	95.97	97.09	97.19	94.69
por.rst.cstn	94.06	93.98	94.35	93.50	94.36	94.11
rus.rst.rst	86.80	85.58	86.21	84.75	85.41	85.74
spa.rst.rststb	96.19	93.53	92.22	96.32	93.70	91.76
spa.rst.sctb	86.88	85.63	82.48	85.44	84.21	80.86
zho.rst.gcdt	92.69	92.55	-	92.20	91.74	-
zho.rst.sctb	79.05	81.84	83.34	77.53	78.55	76.21
eng.sdrst.stac	94.77	95.22	94.91	91.57	90.67	91.91
fra.sdrst.annodis	90.27	88.21	90.02	90.17	90.89	85.78
*eng.dep.covdtb	91.32	92.13	-	91.65	92.13	-
eng.dep.scidtb	96.18	95.07	-	95.63	94.49	-
zho.dep.scidtb	93.33	89.07	-	93.01	90.04	-
Mean	92.60	91.87	-	92.04	91.43	-
Mean corpora 2021	-	91.77	91.48	-	91.22	89.79
Connective identification						
eng.pdtb.pdtb	94.41	93.66	92.02	93.94	91.64	92.56
*eng.pdtb.tedtm	75.86	78.36	-	80.00	75.83	-
ita.pdtb.luna	79.72	65.85	-	74.19	71.60	-
por.pdtb.crpc	85.16	80.66	-	84.65	79.49	-
*por.pdtb.tedtm	73.08	80.29	-	71.22	79.45	-
*tha.pdtb.tdtb	87.43	85.66	-	74.32	69.92	-
tur.pdtb.tdb	89.73	92.77	94.11	89.69	91.12	93.56
*tur.pdtb.tedtm	65.42	64.10	-	64.15	64.78	-
zho.pdtb.cdtb	87.66	89.00	87.52	87.77	90.38	88.35
Mean	82.05	81.15	-	79.99	79.36	-
Mean corpora 2021	-	91.81	91.22	-	91.05	91.49

2023 : PLM multilingue segmentation + connecteurs

Tester simplement un gros modèle multilingue (architecture Transformer) : **xlm-RoBERTa-large**

Problème : c'est un très gros modèle, en fait, impossible à utiliser sur nos serveurs

⇒ on **'freeze'** certains layers = pas d'update

- modèle plus léger
- éventuellement au moins aussi performant : *'lower layers mostly encode morpho-syntactic information, while upper contain more semantic ones'* [Rogers et al., 2020; Kovaleva et al., 2019; Bender and Koller, 2020]

⇒ on **'freeze'** les **6 layers du bas** (on a testé 12 aussi)

Corpus	Treebanked			Plain		
	F1 dev	F1 test	DD21	F1 dev	F1 test	DD21
Segmentation						
deu.rst.pcc	96.79	96.01	95.58	96.60	94.24	93.94
eng.rst.gum	95.54	95.50	94.15	95.78	94.46	92.61
eng.rst.rstdt	97.33	97.62	96.64	97.60	97.74	96.35
eus.rst.ert	91.69	89.93	90.46	91.83	91.09	90.47
fas.rst.prstc	93.79	93.40	92.94	94.05	93.36	92.86
nld.rst.nldt	97.51	96.54	95.97	97.09	97.19	94.69
por.rst.cstn	94.06	93.98	94.35	93.50	94.36	94.11
rus.rst.rrt	86.80	85.58	86.21	84.75	85.41	85.74
spa.rst.rststb	96.19	93.53	92.22	96.32	93.70	91.76
spa.rst.sctb	86.88	85.63	82.48	85.44	84.21	80.86
zho.rst.gcdt	92.69	92.55	-	92.20	91.74	-
zho.rst.sctb	79.05	81.84	83.34	77.53	78.55	76.21
eng.sdr.tstac	94.77	95.22	94.91	91.57	90.67	91.91
fra.sdr.tannodis	90.27	88.21	90.02	90.17	90.89	85.78
*eng.dep.covdtb	91.32	92.13	-	91.65	92.13	-
eng.dep.scidtb	96.18	95.07	-	95.63	94.49	-
zho.dep.scidtb	93.33	89.07	-	93.01	90.04	-
Mean	92.60	91.87	-	92.04	91.43	-
Mean corpora 2021	-	91.77	91.48	-	91.22	89.79
Connective identification						
eng.pdtb.pdtb	94.41	93.66	92.02	93.94	91.64	92.56
*eng.pdtb.tedm	75.86	78.36	-	80.00	75.83	-
ita.pdtb.luna	79.72	65.85	-	74.19	71.60	-
por.pdtb.crpc	85.16	80.66	-	84.65	79.49	-
*por.pdtb.tedm	73.08	80.29	-	71.22	79.45	-
*tha.pdtb.tdtb	87.43	85.66	-	74.32	69.92	-
tur.pdtb.tdb	89.73	92.77	94.11	89.69	91.12	93.56
*tur.pdtb.tedm	65.42	64.10	-	64.15	64.78	-
zho.pdtb.cdtb	87.66	89.00	87.52	87.77	90.38	88.35
Mean	82.05	81.15	-	79.99	79.36	-
Mean corpora 2021	-	91.81	91.22	-	91.05	91.49

2023 : PLM multilingue segmentation + connecteurs

Tester simplement un gros modèle multilingue (architecture Transformer) : **xlm-RoBERTa-large**

Résultats de segmentation : 91.87 / 91.43

- légèrement meilleur que le gagnant de 2021 sur les corpus commun
- score > 85 pour la majorité des corpus (sauf zho.rst.sctb) et proche de 90 pour la majorité:
 - rus.rst.rrt : documents scientifiques contenant la bibliographie qui est une grande EDU + insertion probable de nom d'image segmenté bizarrement
 - spa.rst.sctb : comme la version chinoise, petit corpus

Corpus	Treebanked			Plain		
	F1 dev	F1 test	DD21	F1 dev	F1 test	DD21
Segmentation						
deu.rst.pcc	96.79	96.01	95.58	96.60	94.24	93.94
eng.rst.gum	95.54	95.50	94.15	95.78	94.46	92.61
eng.rst.rstdt	97.33	97.62	96.64	97.60	97.74	96.35
eus.rst.ert	91.69	89.93	90.46	91.83	91.09	90.47
fas.rst.prstc	93.79	93.40	92.94	94.05	93.36	92.86
nld.rst.nldt	97.51	96.54	95.97	97.09	97.19	94.69
por.rst.cstn	94.06	93.98	94.35	93.50	94.36	94.11
rus.rst.rrt	86.80	85.58	86.21	84.75	85.41	85.74
spa.rst.rststb	96.19	93.53	92.22	96.32	93.70	91.76
spa.rst.sctb	86.88	85.63	82.48	85.44	84.21	80.86
zho.rst.gcdt	92.69	92.55	-	92.20	91.74	-
zho.rst.sctb	79.05	81.84	83.34	77.53	78.55	76.21
eng.sdr.tstac	94.77	95.22	94.91	91.57	90.67	91.91
fra.sdr.tannodis	90.27	88.21	90.02	90.17	90.89	85.78
*eng.dep.covdtb	91.32	92.13	-	91.65	92.13	-
eng.dep.scidtb	96.18	95.07	-	95.63	94.49	-
zho.dep.scidtb	93.33	89.07	-	93.01	90.04	-
Mean	92.60	91.87	-	92.04	91.43	-
Mean corpora 2021	-	91.77	91.48	-	91.22	89.79
Connective identification						
eng.pdtb.pdtb	94.41	93.66	92.02	93.94	91.64	92.56
*eng.pdtb.tedm	75.86	78.36	-	80.00	75.83	-
ita.pdtb.luna	79.72	65.85	-	74.19	71.60	-
por.pdtb.crpc	85.16	80.66	-	84.65	79.49	-
*por.pdtb.tedm	73.08	80.29	-	71.22	79.45	-
*tha.pdtb.tdtb	87.43	85.66	-	74.32	69.92	-
tur.pdtb.tdb	89.73	92.77	94.11	89.69	91.12	93.56
*tur.pdtb.tedm	65.42	64.10	-	64.15	64.78	-
zho.pdtb.cdtb	87.66	89.00	87.52	87.77	90.38	88.35
Mean	82.05	81.15	-	79.99	79.36	-
Mean corpora 2021	-	91.81	91.22	-	91.05	91.49

2023 : PLM multilingue segmentation + connecteurs

Tester simplement un gros modèle multilingue (architecture Transformer) : **xlm-RoBERTa-large**

Résultats pour les connecteurs : 81.15 / 79.36

- idem : différence très faible par rapport au gagnant de 2021
- **mais** différence très nette entre les scores de 2021 et ceux de 2023 80-81 vs 91-92 !
 - ajout de nombreux corpus
 - notamment OOD = TED
 - et corpus 'difficile' = LUNA : speech
 - (Thai : erreur dans les données, nouveaux scores autour de 90)

Corpus	Treebanked			Plain		
	F1 dev	F1 test	DD21	F1 dev	F1 test	DD21
Segmentation						
deu.rst.pcc	96.79	96.01	95.58	96.60	94.24	93.94
eng.rst.gum	95.54	95.50	94.15	95.78	94.46	92.61
eng.rst.rstdt	97.33	97.62	96.64	97.60	97.74	96.35
eus.rst.ert	91.69	89.93	90.46	91.83	91.09	90.47
fas.rst.prstc	93.79	93.40	92.94	94.05	93.36	92.86
nld.rst.nldt	97.51	96.54	95.97	97.09	97.19	94.69
por.rst.cstn	94.06	93.98	94.35	93.50	94.36	94.11
rus.rst.rrt	86.80	85.58	86.21	84.75	85.41	85.74
spa.rst.rststb	96.19	93.53	92.22	96.32	93.70	91.76
spa.rst.sctb	86.88	85.63	82.48	85.44	84.21	80.86
zho.rst.gcdt	92.69	92.55	-	92.20	91.74	-
zho.rst.sctb	79.05	81.84	83.34	77.53	78.55	76.21
eng.sdr.tstac	94.77	95.22	94.91	91.57	90.67	91.91
fra.sdr.tannodis	90.27	88.21	90.02	90.17	90.89	85.78
*eng.dep.covdtb	91.32	92.13	-	91.65	92.13	-
eng.dep.scidtb	96.18	95.07	-	95.63	94.49	-
zho.dep.scidtb	93.33	89.07	-	93.01	90.04	-
Mean	92.60	91.87	-	92.04	91.43	-
Mean corpora 2021	-	91.77	91.48	-	91.22	89.79
Connective identification						
eng.pdtb.pdtb	94.41	93.66	92.02	93.94	91.64	92.56
*eng.pdtb.ted	75.86	78.36	-	80.00	75.83	-
ita.pdtb.luna	79.72	65.85	-	74.19	71.60	-
por.pdtb.crpc	85.16	80.66	-	84.65	79.49	-
*por.pdtb.ted	73.08	80.29	-	71.22	79.45	-
*tha.pdtb.tdtb	87.43	85.66	-	74.32	69.92	-
tur.pdtb.tdb	89.73	92.77	94.11	89.69	91.12	93.56
*tur.pdtb.ted	65.42	64.10	-	64.15	64.78	-
zho.pdtb.cdtb	87.66	89.00	87.52	87.77	90.38	88.35
Mean	82.05	81.15	-	79.99	79.36	-
Mean corpora 2021	-	91.81	91.22	-	91.05	91.49

Bilan segmentation et connecteurs

- Puissance des modèles Transformers et embeddings contextuels

Segmentation :

- Performance moyenne au-dessus de 90% : plateau → analyse d'erreurs pour comprendre ce qui est encore difficile: informations manquantes au niveau des embeddings ? erreurs d'annotations / règles trop complexes ?
- Scores similaires quand on n'a pas les frontières de phrases

Connecteurs : beaucoup de nouveaux corpus en 2023

- Scores bas pour des genres spécifiques (speech) et données out-of-domain (cf TED)

Identification des relations

- Implicites : considérée comme une tâche très difficile car :
 - complexe : informations à plein de niveaux (lexique, syntaxe, sémantique, connaissances du monde) + interactions entre les paires de segments / avec le reste du document
 - 'peu' de données (environ 18 000 annotations dans le PDTB)
 - surtout si on considère toutes les relations (distributions très déséquilibrée)
- Explicites : considérée comme résolue, mais surtout étudiée pour l'anglais et les articles de journaux

DISRPT : peu de propositions profitant du cadre multilingue

DiscReT : but = proposer un système unique quelque soit la langue et profiter de l'ensemble des données

DiscReT: Discourse Relation Tagging

- (1) Approche multilingue : utilisable pour n'importe quelle langue
- (2) Approche zero-shot : performance sur une langue inconnue à l'entraînement ?
 - basé sur architecture Transformer (BERTM ou xlm-RoBERTa)
 - **entraînement joint sur l'ensemble des données ou des sous-groupes**

[Meteniti et al 2023; COLING under review]

DiscReT: Discourse Relation Tagging

- (1) Approche multilingue : utilisable pour n'importe quelle langue
- (2) Approche zero-shot : performance sur une langue inconnue à l'entraînement ?
 - basé sur architecture Transformer (BERTM ou xlm-RoBERTa)
 - **entraînement joint sur l'ensemble des données ou des sous-groupes**
 - par groupe de langues
 - par framework
 - par similarité de jeux de relations = similarité Jaccard : des corpus sont proches quand l'intersection de leurs ensembles de relations est large

[Meteniti et al 2023; COLING under review]

DiscReT: Discourse Relation Tagging

- (1) Approche multilingue : utilisable pour n'importe quelle langue
- (2) Approche zero-shot : performance sur une langue inconnue à l'entraînement ?
 - basé sur architecture Transformer (BERTM ou xlm-RoBERTa)
 - **entraînement joint sur l'ensemble des données ou des sous-groupes**
 - par groupe de langues
 - par framework
 - par **similarité de jeux de relations** = similarité Jaccard : des corpus sont proches quand l'intersection de leurs ensembles de relations est large

[Meteniti et al 2023; COLING under review]

Jaccard

corpus PDTB dont le jeu de relation est totalement différent des autres

zho.pdtb.cdtb

eng.rst.gum

corpus RST dont le jeu de relation est totalement différent des autres

zho.rst.gcdt

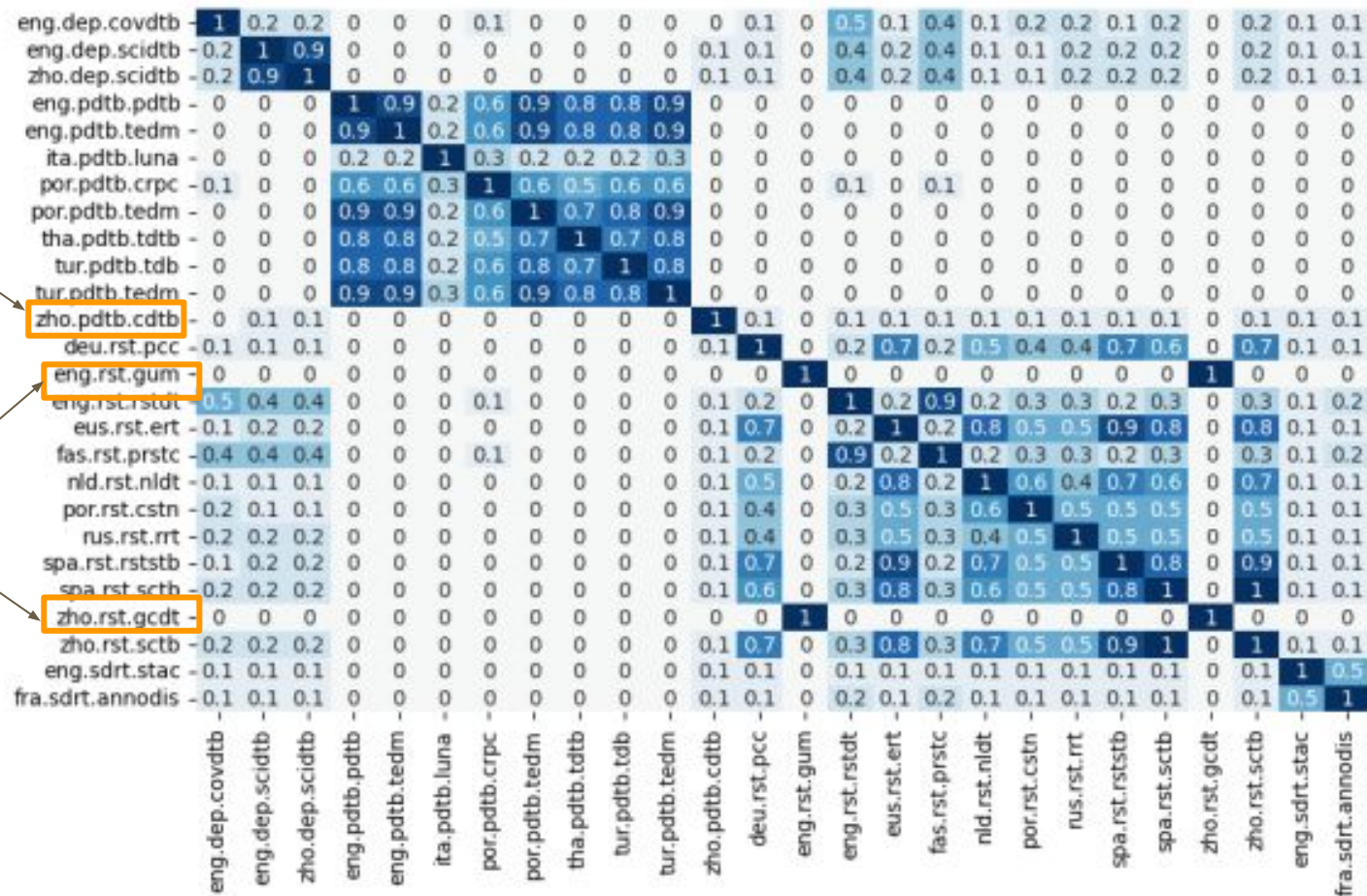


Figure 3.5: Jaccard similarities heatmap.

DiscReT: Discourse Relation Tagging

Problème : **164 labels différents!** Et des jeux différents pour chaque corpus

- Homogénéisation additionnelle → 135 labels, toujours beaucoup ...
- **Filtrage sur les labels :**
 - avec un entraînement sur des corpus différents, beaucoup de confusion
 - → on force le modèle à prédire des relations du framework cible
 - bruité mais plus précis

L'ajout de traits + le filtrage ⇒ un système multilingue entraîné sur toutes les données obtient des performances similaires à un système monolingue

- pas évident car les différents corpus amènent beaucoup de bruit

Résultats

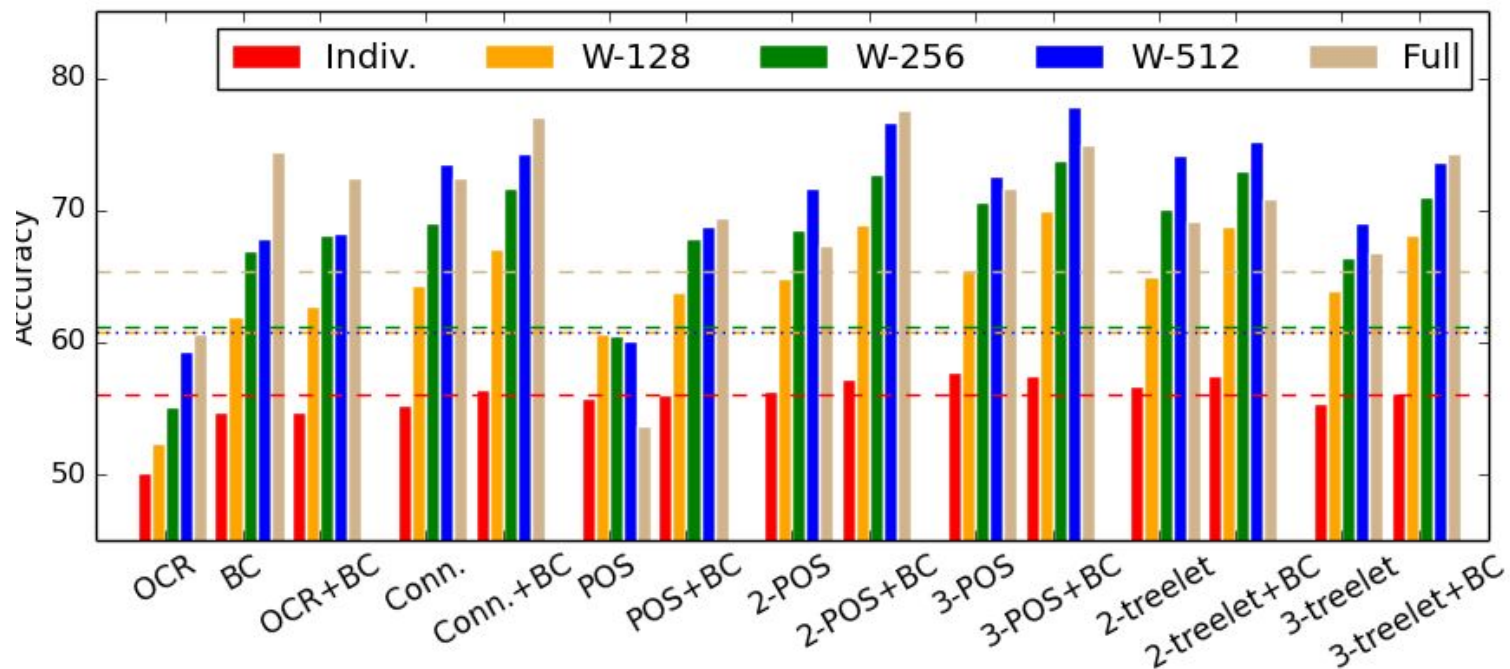
- scores en 0-shot en général bas / monolingue → mais pas catastrophiques
- en général : meilleur avec Jaccard → la divergence en termes de jeux de relations est cruciale

corpus	monolingual	0-shot - framework	0-shot - Jaccard
deu.rst.pcc	32	20	18
eus.rst.ert	42	33	36
nld.rst.nldt	43	30	31
por.rst.cstn	57	49	46
rus.rst.rrt	59	40	31
spa.rst.rststb	56	46	49
spa.rst.sctb	43	60	61
zho.rst.sctb	46	48	51
eng.dep.scidtb	72	35	40
eng.rst.rstdt	64	42	37
fas.rst.prstc	52	40	46
zho.dep.scidtb	55	41	43

Evaluation applicative

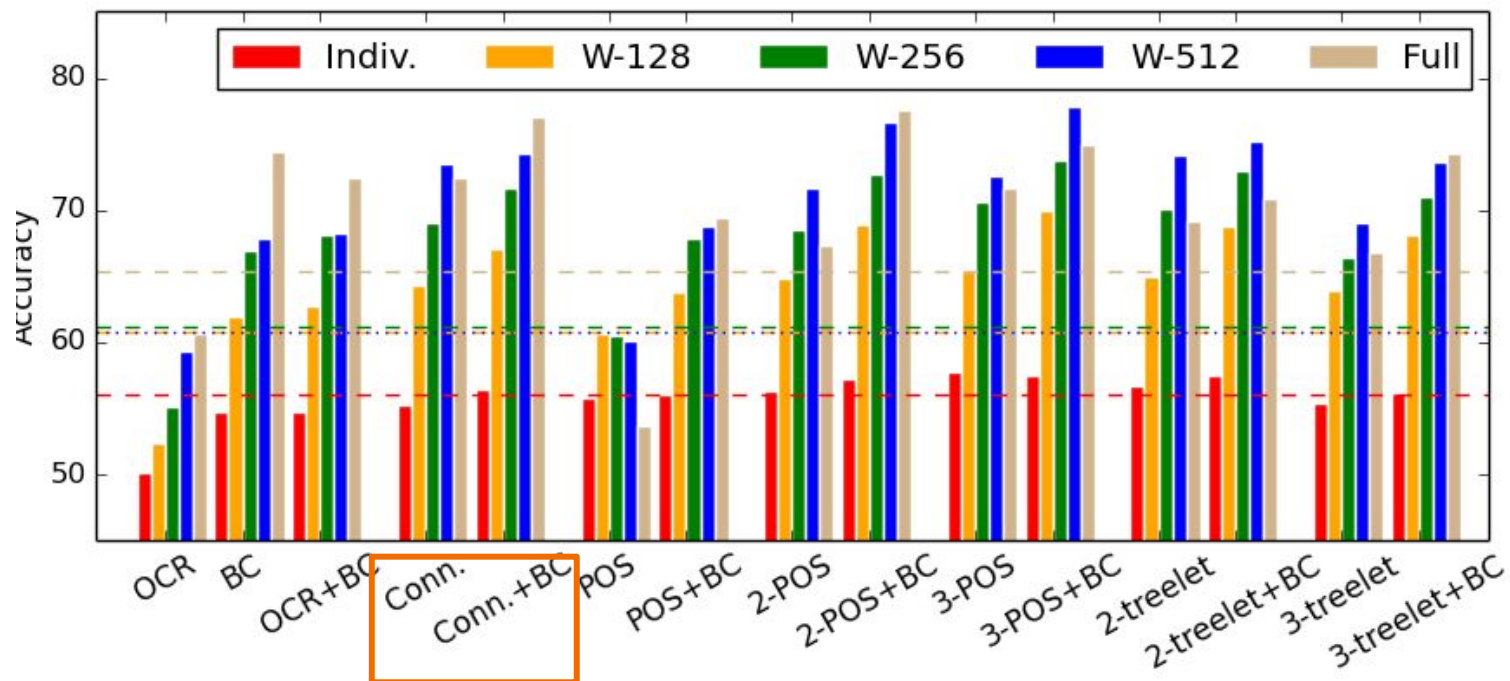
Utiliser les connecteurs

Identification des marqueurs langagiers de la schizophrénie



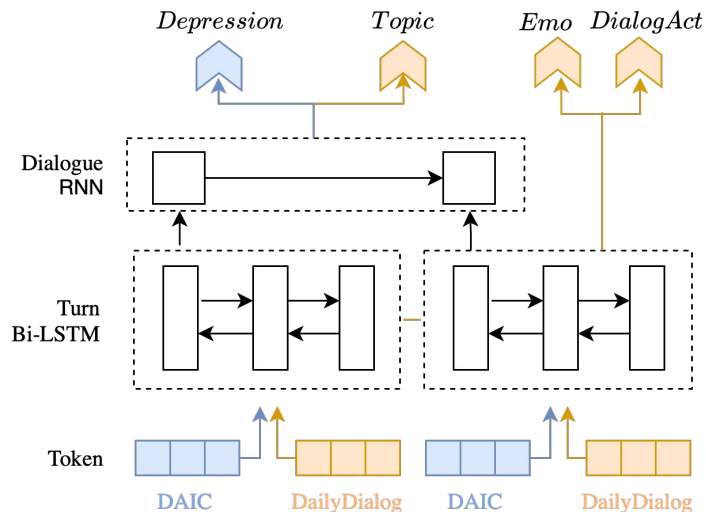
Utiliser les connecteurs

Identification des marqueurs langagiers de la schizophrénie



Apprentissage multi-tâche [Li et al 2022]

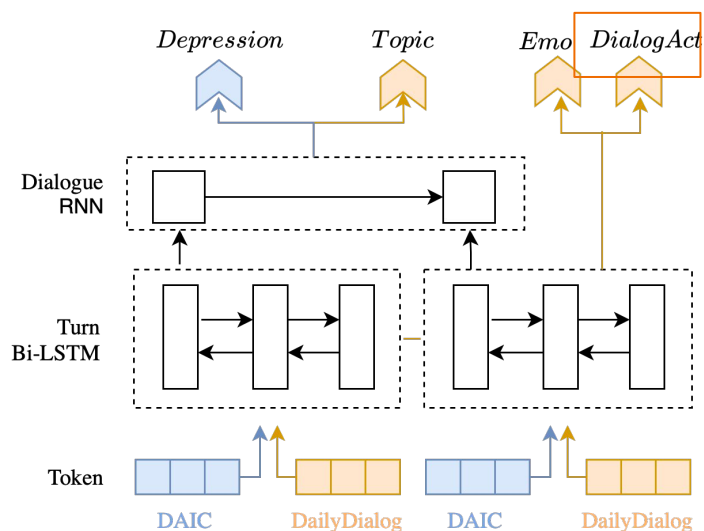
Identification de personnes souffrant de dépression (DAIC + DailyDialog)



	F ₁	Prec.	Rec.	Acc.
BSL Majority vote	41.3	35.1	50.0	70.2
<i>State-of-the-art</i>				
NHN (baseline) [5]	45	-	50	-
HCAN [5]	63	-	66	-
HAN+L [6]	70	-	70	-
<i>Ours</i>				
STL Depression	43.9	44.5	47.5	63.8
MTL +Emo	55.5	56.2	61.6	70.2
MTL +Top	55.6	55.9	56.8	59.6
MTL +Diag	60.8	60.6	61.4	66.0
MTL +Emo+Diag+Top	70.6*	70.1	71.5*	74.5

Apprentissage multi-tâche [Li et al 2022]

Identification de personnes souffrant de dépression (DAIC + DailyDialog)



	F ₁	Prec.	Rec.	Acc.
BSL Majority vote	41.3	35.1	50.0	70.2
<i>State-of-the-art</i>				
NHN (baseline) [5]	45	-	50	-
HCAN [5]	63	-	66	-
HAN+L [6]	70	-	70	-
<i>Ours</i>				
STL Depression	43.9	44.5	47.5	63.8
MTL +Emo	55.5	56.2	61.6	70.2
MTL +Top	55.6	55.9	56.8	59.6
MTL +Diag	60.8	60.6	61.4	66.0
MTL +Emo+Diag+Top	70.6*	70.1	71.5*	74.5

Structures discursives latentes [Devatine et al 2022]

Etude de biais politiques mais sans analyseurs discursifs, via des structures latentes (mais basé sur découpage en EDUs vs phrases)

"The Virginia Beach shooter put a sound suppressor (...) so that the death shots were muffled, perhaps denying others the warning that would have allowed them to escape. It is long past time to remove the silencer that seems to suppress action on gun-control legislation, to treat mass shooting as the epidemic it is, and do everything possible to save lives."
(Washington Post, left-leaning)

"The attack began shortly after 16:00 (20:00 GMT), at Virginia Beach Municipal Center, in an area which is home to a number of city government buildings. The area was put into lockdown by police and employees were evacuated. 'We just heard people yelling and screaming at people to get down,' Megan Banton, an administrative assistant in the building, told local television news station WAVY."
(BBC, center)

"The chilling fact is that mass public killers are attracted to targets where people can't defend themselves. (...) Ninety-eight percent of US mass public shootings since 1950 have occurred in places where people weren't allowed to defend themselves. But the news media refuses to cover this fact, which illustrates the need for self-defense, not for more gun control that doesn't work."
(Townhall, right-leaning)

Model	Acc.	Macro F_1	MAE
Ours Base	46.97	44.41	0.69
Ours+SA/Sent	48.76	45.84	0.67
Ours+SA/EDU	54.39	51.36	0.57
Ours: 512t, +SA/EDU	50.04	45.23	0.70
Baly 20: 512t, LSTM	46.42	45.44	0.62
Baly 20: 512t, BERT	51.41	48.26	0.51

Conclusion

Défis actuels

- Toujours gérer le manque de données (annotation très coûteuse)
 - [Carrenini et al 2019] : détection de sentiment pour inférer la structure discursive d'un document (distant learning) → résultat mitigé (structure + nucléarité)
 - [Liu and Lapata, 2017; Karimi et al, 2019] : structures latentes (pas besoin de parser)
 - Travaux en crowd-sourcing [], co-training []
- Note : on a quand même besoin d'annotations :)
 - idéalement en variant langues / domaines / modalités → nouvelles données en anglais
- Toujours sortir de l'anglais, du WSJ et des monologues
 - [Zeldes et al 2023] : sur les questions de genre de documents
- Nécessité d'une réflexion linguistique
 - Différences entre formalismes, langues / domaines / modalités ? Unification des jeux de relations ?
 - Quelles informations sont nécessaires pour l'analyse discursive ? comment les intégrer ?
- Evaluation applicative
 - évaluation extrinsèque nécessaire
 - Comment utiliser l'information discursive pour des applications ?

Champ de recherche en plein essor

- De plus en plus de publications dans les conférences, surtout :
 - identification des relations implicites
 - parsing RST
 - léger regain pour la segmentation
- Workshops dédiés :
 - LSDsem, DiscoMT, DisRPT, DSSNLG
 - CoDi in 2020, 2021, 2022, 2023 et 2024 ! ⇒ **deadline le 20/12** (avec EACL, Malta)
- Intérêt industriel :
 - quantum (anr) : prise en compte de la structure des documents pour la génération de question (amélioration de FAQs)
 - summ-re (anr) : résumé automatique de réunions

Données de segmentation

<https://github.com/disrpt/sharedtask2023>

Segmenteur “utilisable” pour le Français (nouvelle version en cours de développement):

<https://gitlab.irit.fr/melodi/andiamo/discoursesegmentation/discut>

Merci !

*And even in our wildest and most wandering reveries, nay in our very dreams, we shall find, if we reflect, that **the imagination ran not altogether at adventures**, but that there was still a connection upheld among the different ideas, which succeeded each other. Were the loosest and freest conversation to be transcribed, there would immediately be transcribed, there would immediately be observed **something which connected it in all its transitions**.*

Cross-lingual RST discourse parsing

- Systèmes **monolingues si > 100 arbres**,
- Sinon combiner les corpus en utilisant ou non des données cibles en validation
- Différences dans les schémas d'annotation = harmonisation
- Parsers en constituents avec embeddings pré-entraînés et beam search

Corpus	#Arbres	#Mots	#Rel	#EDUs
EN	385	206 300	56	21 789
PT	329	135 820	32	12 573
ES	266	69 787	29	4 019
DE	173	32 274	30	2 790
NL	80	27 920	31	2 345
EU	85	27 982	31	2 396

Résultats [Braud et al. 2017]

System	En-DT			Pt-DT			Es-DT			De-DT			NI-DT			Eu-DT		
	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel
MFS	58.2	33.4	22.1	57.3	33.9	23.23	82.0	51.5	17.7	61.3	37.8	13.2	57.9	35.5	22.0	63.2	34.9	18.8
Li et al.	85.0	70.8	58.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ji and Eisenstein	82.1	71.1	61.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mono	85.0	72.3	60.1	82.0	65.1	49.9	89.7	72.7	54.4	80.2	53.9	35.0	-	-	-	-	-	-
+ emb.	83.5	68.5	55.9	81.3	62.9	48.8	89.3	72.4	51.4	77.7	51.6	31.1	-	-	-	-	-	-
Source only	76.3	50.5	31.3	76.5	54.6	35.5	78.1	45.4	27.0	76.0	46.0	26.1	69.5	42.1	25.3	78.6	53.0	26.4
Source+Target	85.1	73.1	61.4	81.9	65.1	49.8	88.8	68.0	50.4	79.6	53.6	34.1	69.2	43.4	28.3	76.7	50.5	29.5

Source + Target = train on Pt+Es+De+En ; test on En

- improvement for relations largely represented in all corpora (e.g. JOINT +3%),
- or under-represented in the EnDT (e.g. CONDITION +3%).

Matrice de confusion (Basque)

RST relation	Match	
ELABORATION	101	0.616
SAME-UNIT	40	0.244
EVALUATION	9	0.055
BACKGROUND	6	0.036
MEANS	3	0.018
CAUSE	2	0.012
ENABLEMENT	2	0.012
JOINT	1	0.006
Total agreement	164	

Relation	Errors	Empl. Tags
ELABORATION	213	314
BACKGROUND	252	258
JOINT	191	192
CAUSE	77	79
SAME-UNIT	67	107
EVALUATION	63	72
ENABLEMENT	29	31

Table 7: Description of gold and automatic label matching

- machine tries to get the best results using a small number of very general relations such as: ELABORATION, BACKGROUND and JOINT
- disagreement between humans : general, widely used and less informative relations, such as ELABORATION, LIST, BACKGROUND, RESULT and MEANS

Table 8: Parser annotation confusion matrix

Finalemment :

- difficile de combiner des corpus : problème de langues / adaptation ? ou de différences dans les schémas d'annotations ?
- + problème d'évaluation : un système qui prédit surtout bien des élaborations n'est peut-être pas très utile ...

(RST / SDRT) Discourse Parsing

- Parsing discursif ≈ parsing syntaxique : arbres couvrant un document
 - en général, mêmes méthodes : shift-reduce, CKY, format dépendance...
- Difficultés spécifiques :
 - combinatoire / efficacité
 - représentation des données
 - manque de données (385 arbres RST)
 - + manque de rigueur dans l'évaluation, cf [Morey et al 2016]
- Beaucoup de travaux :
 - apprendre une représentation [Ji and Eisenstein,], [Multi-view and multi-task training of RST discourse parsers](#) (2016)
 - diviser la tâche : arbres nus puis classifieurs pour les relations [Wang et al, 2017]
 - études cross-lingues : [Cross-lingual RST discourse parsing](#) (2017), [EusDisParser](#) (2019), [Liu et al. 2020]
 - attachement en supervision distante, cf [Huber et Carenini], travaux en cours de Chuyuan Li

Multi-task and multi-view training of RST parsers

- Motivation: manque de données et beaucoup d'info nécessaire
- Solution: récupérer de l'information via d'autres données

→ **Apprentissage multi-tâche pour combiner différents corpus**

- Corpus discursifs suivant des frameworks différents (RST DT et PDTB)
- Données pour d'autres tâches liées : temps, factualité, coréférence, segmentation en tours de paroles...
- Différentes représentations de la sortie, i.e. dépendance / constituance, granularité des étiquettes

→ Réseau de neurones avec 'hard-parameter sharing' (partage des couches cachées, une couche de sortie par tâche)

Les 'vues' et les 'tâches'

Alternate views

	Constituent (main)	Dependency	Nuclearity	Relation	Fine-grained
EDU 1	(NN-TextOrg(NN-SaUnit(NN-List	Root	(NN(NN(NN	(TextOrg(SaUnit(List	(TextOrg(SaUnit(List
EDU 2	NN-List)	-1 NN-List	NN)	List)	List)
EDU 3	(NS-Elab	-2 NN-SaUnit	(NS	(Elab	(Elab-set-member

Auxiliary tasks

	Speech	Factuality	Aspect	Modality	Polarity	Tense	Coreference	PDTB
Corpus	Santa Barbara	Factbank	Timebank	Timebank	Timebank	Timebank	Ontonotes	PDTB
Sent 1	turn1	Certain	Progressive	Must	Positive	Past	Root	Root
Sent 2	turn2	Probable	Perfective	Could	Negative	Future	Coreferent	Contrast

Résultats

System	Fine	Fact	Speech	Asp	Dep	Nuc+lab	Mod	Pol	PDTB	Coref	Ten	Span	Nuclearity	Relation
Prior work														
DPLP concat	-	-	-	-	-	-	-	-	-	-	-	82.08	71.13	61.63
DPLP general	-	-	-	-	-	-	-	-	-	-	-	81.60	70.95	61.75
Feng and Hirst [2014]	-	-	-	-	-	-	-	-	-	-	-	84.9	69.9	57.2
Feng and Hirst [2014] PE	-	-	-	-	-	-	-	-	-	-	-	85.7	71.0	58.2
Our work														
Hier-LSTM	-	-	-	-	-	-	-	-	-	-	-	81.39	64.54	49.15
MTL-Hier-LSTM	✓	-	-	-	-	-	-	-	-	-	-	82.88	67.46	53.25
MTL-Hier-LSTM	-	✓	-	-	-	-	-	-	-	-	-	83.40	67.16	52.10
MTL-Hier-LSTM	-	-	✓	-	-	-	-	-	-	-	-	83.26	67.51	51.75
MTL-Hier-LSTM	-	-	-	✓	-	-	-	-	-	-	-	83.69	66.25	51.25
MTL-Hier-LSTM	-	-	-	-	✓	-	-	-	-	-	-	81.25	65.34	51.24
MTL-Hier-LSTM	-	-	-	-	-	✓	-	-	-	-	-	82.09	65.68	51.12
MTL-Hier-LSTM	-	-	-	-	-	-	✓	-	-	-	-	81.66	65.31	50.58
MTL-Hier-LSTM	-	-	-	-	-	-	-	✓	-	-	-	82.01	65.29	50.11
MTL-Hier-LSTM	-	-	-	-	-	-	-	-	✓	-	-	81.61	63.10	48.89
MTL-Hier-LSTM	-	-	-	-	-	-	-	-	-	✓	-	80.26	63.35	47.70
MTL-Hier-LSTM	-	-	-	-	-	-	-	-	-	-	✓	81.33	62.34	47.57
Best combination	-	-	-	-	✓	✓	✓	-	✓	-	-	83.62	69.77	55.11
Human annotation	-	-	-	-	-	-	-	-	-	-	-	88.70	77.72	65.75

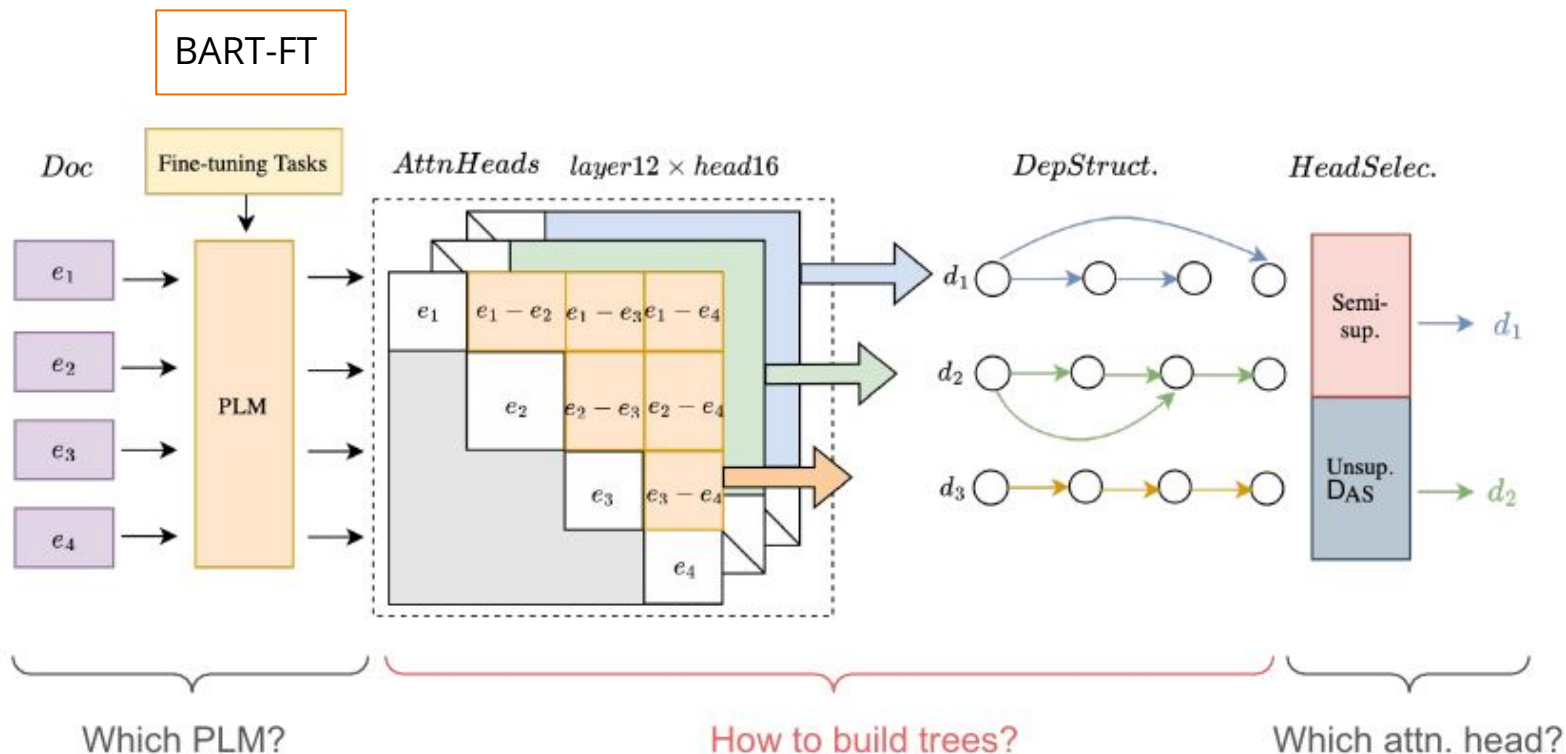
- MTL améliore par rapport à STL pour 8/11tasks
- Les différentes représentations de la sortie aident beaucoup
- Factualité et Tours de parole aident beaucoup
- Temps et Coréférence n'aident pas, sûrement un mauvais encodage

Weakly supervised structure prediction [Chuyuan Li et al 2023?]

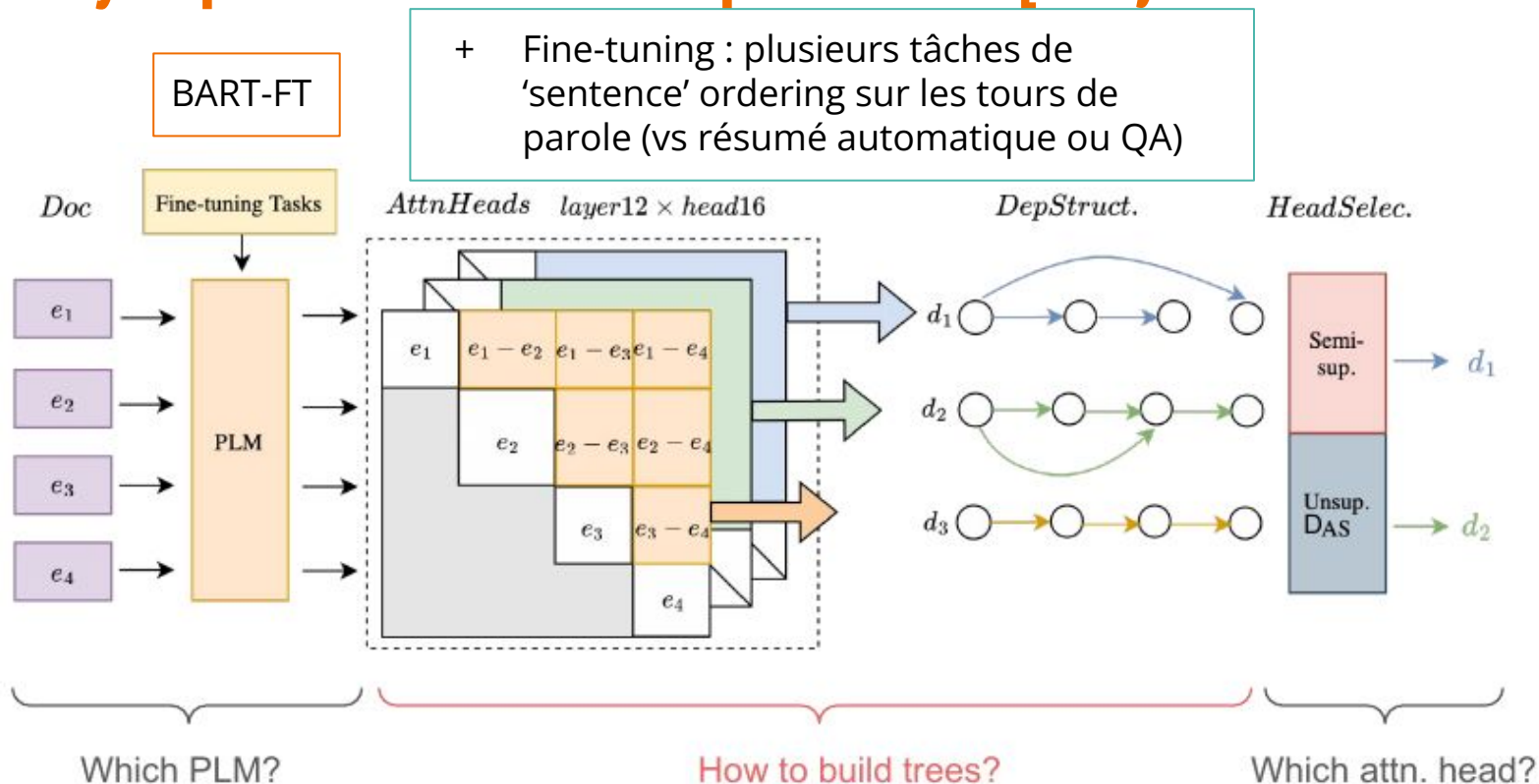
BERTology: tâches d'extraction de structure discursive dans des modèles de langues pré-entraînés (PLMs): [Koto et al., 2021, Pandia et al.. 2021, Huber&Carenini 2022] → ici extraction de la structure des dialogues

- Idée : fine-tuner un PLM sur une tâche liée (pas de supervision directe)
- Utiliser les scores d'attention pour inférer un attachement entre EDUs : matrice d'attention sur les tokens → matrice sur les EDUs → distance
- Comparer les structures obtenues à des structures gold (STAC)

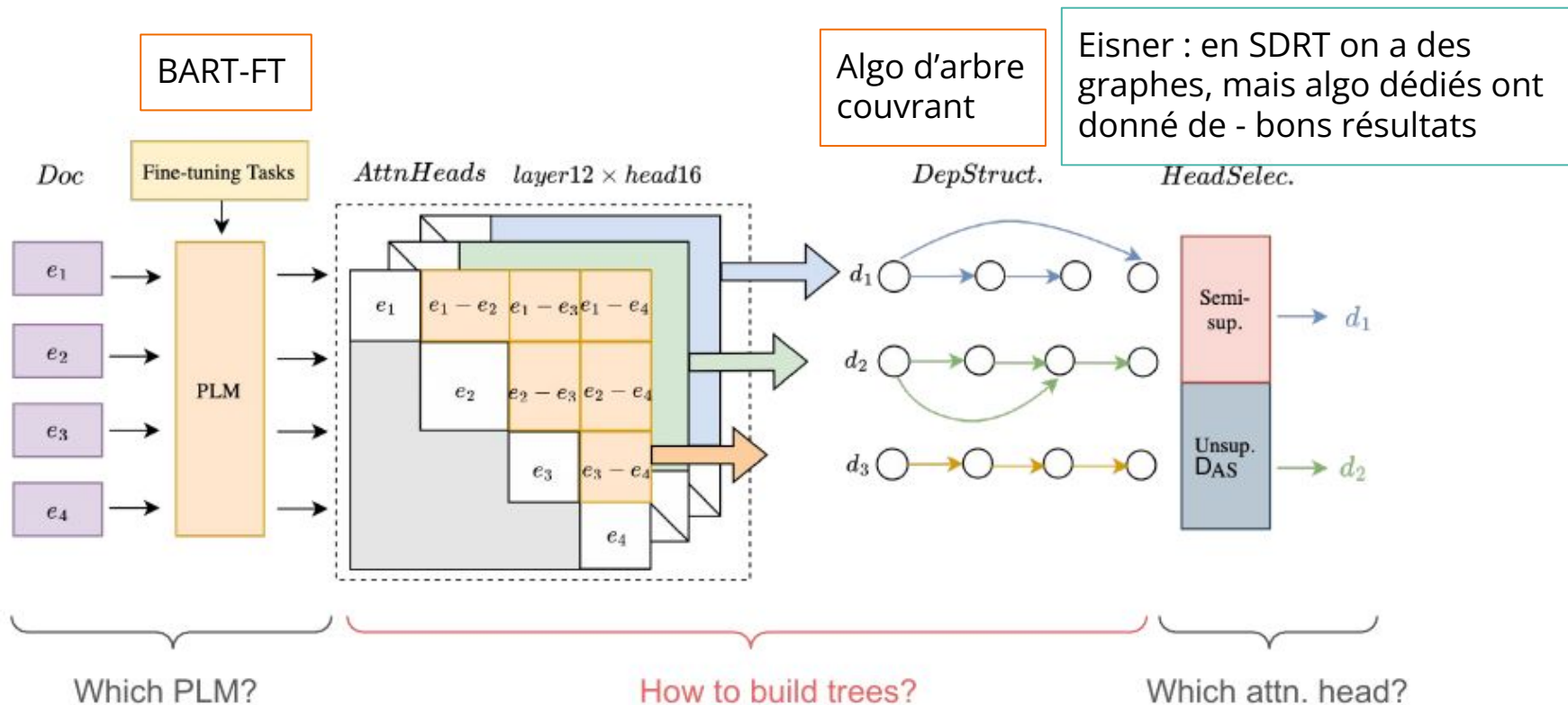
Weakly supervised structure prediction [Chuyuan Li et al 2023?]



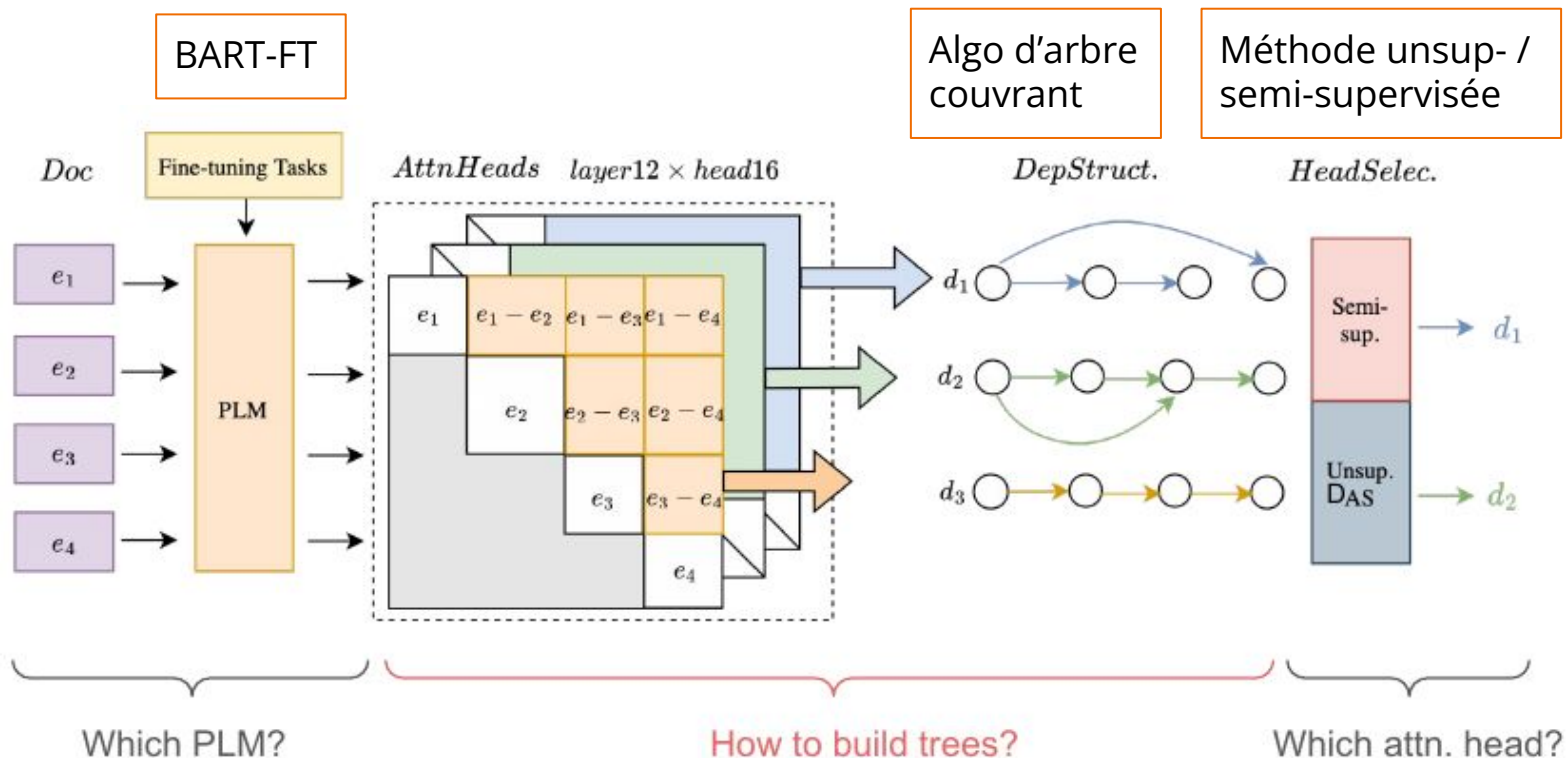
Weakly supervised structure prediction [Chuyuan Li et al 2023?]



Weakly supervised structure prediction [Chuyuan Li et al 2023?]



Weakly supervised structure prediction [Chuyuan Li et al 2023?]



Weakly supervised structure prediction [Chuyuan Li et al 2023?]

Model			
<i>Unsupervised Baseline</i>			
LAST	56.8		
<i>Supervised Models</i>			
Deep-Sequential shi2019deep	71.4		
SSA-GNN wangstructure2021	73.8		
<i>Unsupervised PLMs</i>			
	H _g	H _l	H _{ora}
BART	56.6	56.4	57.6
+ CNN	56.8	56.7	57.1
+ SAMSum	56.7	56.6	57.6
+ SQuAd2	55.9	56.4	57.7
+ SO-DD	56.8	57.1	58.2
+ SO-STAC	56.7	57.2	59.5

(a) Micro-F₁ on STAC for supervised SOTA models and PLMs. H_g: global best head. H_l: local best heads. H_{ora}: oracle head. Best (non-oracle) score in the 3rd block in bold.

Train on →	BART	+ SO-DD	+ SO-STAC
Test with ↓	F ₁	F ₁	F ₁
LAST BSL	56.8	56.8	56.8
Gold H	57.6	58.2	59.5
Unsup H _g	56.6	56.8	56.7
Unsup H _l	56.4	57.1	57.2
Few-shot-val 10	57.0 _{0.012}	57.2 _{0.012}	57.1 _{0.026}
Few-shot-val 30	57.3 _{0.005}	57.3 _{0.013}	59.2 _{0.009}
Few-shot-val 50	57.4_{0.004}	57.7_{0.005}	59.3_{0.007}

(b) STAC micro-F₁ scores from BART and two fine-tuned models (+SO-DD: BART SO fine-tuned on DD, +SO-STAC: BART fine-tuned on STAC), with unsupervised and semi-supervised approaches. Best score per column (only consider “unsup.” and “few-shot” sections) is in bold. Subscription is standard deviation.

Weakly supervised structure prediction [Chuyuan Li et al 2023?]

Model			
<i>Unsupervised Baseline</i>			
LAST			56.8
<i>Supervised Models</i>			
Deep-Sequential shi2019deep			71.4
SSA-GNN wangstructure2021			73.8
<i>Unsupervised PLMs</i>			
	H _g	H _l	H _{ora}
BART	56.6	56.4	57.6
+ CNN	56.8	56.7	57.1
+ SAMSum	56.7	56.6	57.6
+ SQuAd2	55.9	56.4	57.7
+ SO-DD	56.8	57.1	58.2
+ SO-STAC	56.7	57.2	59.5

(a) Micro-F₁ on STAC for supervised SOTA models and PLMs. H_g: global best head. H_l: local best heads. H_{ora}: oracle head. Best (non-oracle) score in the 3rd block in bold.

Train on → Test with ↓	BART F ₁	+ SO-DD F ₁	+ SO-STAC F ₁
LAST BSL	56.8	56.8	56.8
Gold H	57.6	58.2	59.5
Unsup H _g	56.6	56.8	56.7
Unsup H _l	56.4	57.1	57.2
Few-shot-val 10	57.0 _{0.012}	57.2 _{0.012}	57.1 _{0.026}
Few-shot-val 30	57.3 _{0.005}	57.3 _{0.013}	59.2 _{0.009}
Few-shot-val 50	57.4 _{0.004}	57.7 _{0.005}	59.3 _{0.007}

(b) STAC micro-F₁ scores from BART and two fine-tuned models (+SO-DD: BART SO fine-tuned on DD, +SO-STAC: BART fine-tuned on STAC), with unsupervised and semi-supervised approaches. Best score per column (only consider “unsup.” and “few-shot” sections) is in bold. Subscription is standard deviation.

Weakly supervised structure prediction [Chuyuan Li et al 2023?]

Model			
<i>Unsupervised Baseline</i>			
LAST			56.8
<i>Supervised Models</i>			
Deep-Sequential shi2019deep			71.4
SSA-GNN wangstructure2021			73.8
<i>Unsupervised PLMs</i>			
	H_g	H_l	H_{ora}
BART	56.6	56.4	57.6
+ CNN	56.8	56.7	57.1
+ SAMSum	56.7	56.6	57.6
+ SQuAd2	55.9	56.4	57.7
+ SO-DD	56.8	57.1	58.2
+ SO-STAC	56.7	57.2	59.5

(a) Micro- F_1 on STAC for supervised SOTA models and PLMs. H_g : global best head. H_l : local best heads. H_{ora} : oracle head. Best (non-oracle) score in the 3rd block in bold.

Train on \rightarrow	BART	+ SO-DD	+ SO-STAC
Test with \downarrow	F_1	F_1	F_1
LAST BSL	56.8	56.8	56.8
Gold H	57.6	58.2	59.5
Unsup H_g	56.6	56.8	56.7
Unsup H_l	56.4	57.1	57.2
Few-shot-val 10	57.0 _{0.012}	57.2 _{0.012}	57.1 _{0.026}
Few-shot-val 30	57.3 _{0.005}	57.3 _{0.013}	59.2 _{0.009}
Few-shot-val 50	57.4_{0.004}	57.7_{0.005}	59.3_{0.007}

Non ou
semi-
supervisé

(b) STAC micro- F_1 scores from BART and two fine-tuned models (+SO-DD: BART SO fine-tuned on DD, +SO-STAC: BART fine-tuned on STAC), with unsupervised and semi-supervised approaches. Best score per column (only consider “unsup.” and “few-shot” sections) is in bold. Subscription is standard deviation.

Les corpus

Language	RST	SDRT	PDTB	Other
English	RST DT	DisCor	PDTB	GraphBank
	Instructional	STAC corpus	Biomedical DRB	-
	SFU Review Corpus	-	-	-
English, Spanish, Basque	Multilingual RST DT	-	-	-
Spanish	Spanish RST DT	-	-	-
Italian	-	-	LUNA	-
French	-	ANNODIS	French DT	-
German	PCC	-	(PCC)	Tüba-D/Z
	Twitter?	-	-	-
Dutch	Dutch RST DT	-	-	DiscAn
Danish	-	-	-	Copenhagen DeT
Arabic	-	Arabic DT	Leeds Arabic DT	-
Basque	Basque RST DT	-	-	-
Catalan	-	-	-	CatDiG
Portuguese	Rhetalho	-	-	-
Brazilian	CSTNews	-	-	-
Czech	-	-	Prague DT	-
Turkish	-	-	Turkish DT	-
Chinese	Chinese RST DT	-	Chinese DT	-
Hindi	-	-	Hindi DRB	-
Tamil	?	-	-	-

DT = Discourse Treebank ; DRB = Discourse Relation Bank

SDRT: Molwani: Multiparty Dialogues (English, 2020)

PDTB: TED talks English, Polish, German, Russian, European Portuguese, and Turkish (2019)

SciDTB: Scientific Abstracts (2018) +Turnitin: student essays (2019)

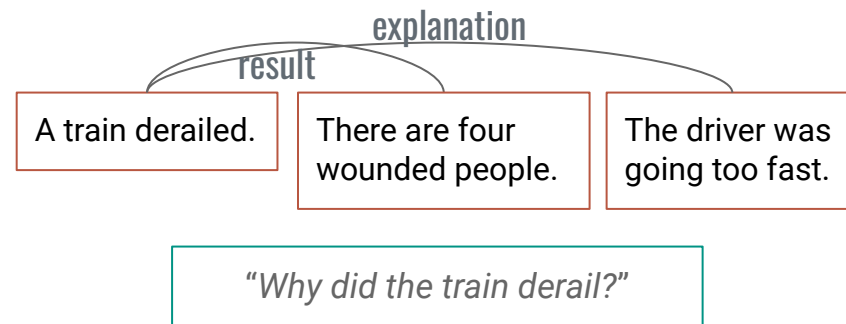
Thème: discourse parsing

→ **Discourse parsing:**

- construire des structure de **documents**
- représentant leur **sens**

→ Permet d'obtenir une analyse fine nécessaire pour certaines applications,

e.g. Question Answering, Summarization, Translation...



Challenges et questions scientifiques

→ Performances basses des systèmes actuels :

- problème d'apprentissage complexe : texte long, structures complexes, sémantique
- peu de données, e.g. 385 instances / arbres pour le corpus le plus utilisé

Challenges et questions scientifiques

→ Performances basses des systèmes actuels :

- problème d'apprentissage complexe : texte long, structures complexes, sémantique
- peu de données, e.g. 385 instances / arbres pour le corpus le plus utilisé

Challenges et questions scientifiques

→ Performances basses des systèmes actuels :

- problème d'apprentissage complexe : texte long, structures complexes, sémantique
- peu de données, e.g. 385 instances / arbres pour le corpus le plus utilisé

what is so hard to learn? Do we have issues with data representation or problem modeling?

Challenges et questions scientifiques

→ Performances basses des systèmes actuels :

- problème d'apprentissage complexe : texte long, structures complexes, sémantique
- peu de données, e.g. 385 instances / arbres pour le corpus le plus utilisé
- ++ portée limitée des systèmes actuels : Anglais et journaux

→ très peu de recherche sur des applications cibles et l'utilité de ces structures

what is so hard to learn? Do we have issues with data representation or problem modeling?

Challenges et questions scientifiques

→ Performances basses des systèmes actuels :

- problème d'apprentissage complexe : texte long, structures complexes, sémantique
- peu de données, e.g. 385 instances / arbres pour le corpus le plus utilisé
- ++ portée limitée des systèmes actuels : Anglais et journaux

→ très peu de recherche sur des applications cibles et l'utilité de ces structures

what is so hard to learn? Do we have issues with data representation or problem modeling?

what do we want to learn? What is needed from discourse analysis for applications?

Challenges et questions scientifiques

→ Performances basses des systèmes actuels :

- problème d'apprentissage complexe : texte long, structures complexes, sémantique
- peu de données, e.g. 385 instances / arbres pour le corpus le plus utilisé
- ++ portée limitée des systèmes actuels : Anglais et journaux

→ très peu de recherche sur des applications cibles et l'utilité de ces structures

→ Plusieurs cadres théoriques (et donc différents corpus) avec des jeux de relations variés et différentes contraintes sur les structures

what is so hard to learn? Do we have issues with data representation or problem modeling?

what do we want to learn? What is needed from discourse analysis for applications?

Challenges et questions scientifiques

→ Performances basses des systèmes actuels :

- problème d'apprentissage complexe : texte long, structures complexes, sémantique
- peu de données, e.g. 385 instances / arbres pour le corpus le plus utilisé
- ++ portée limitée des systèmes actuels : Anglais et journaux

→ très peu de recherche sur des applications cibles et l'utilité de ces structures

→ Plusieurs cadres théoriques (et donc différents corpus) avec des jeux de relations variés et différentes contraintes sur les structures

what is so hard to learn? Do we have issues with data representation or problem modeling?

what do we want to learn? What is needed from discourse analysis for applications?

what did we learn at the end? Could we inform theoretical frameworks?

Methodologie : Multi-objective learning

- différents modèles sont entraînés avec des objectifs différents mais liés (entraînement joint ou non)
- objectif = “tâche”: annotation linguistique (e.g. syntaxe, temps) ou applications (e.g. summarization)

→ Ce cadre d'apprentissage est peu étudié pour une analyse de haut niveau comme le discours

WP1: data representation	W2: Frameworks divergences	W3: Languages and dialogue	WP4: Extrinsic evaluation
M. Coavoux, P. Muller	N. Asher, P. Muller	M. Coavoux, B. Plank	M. Amblard, M. Musiol, V. Gardette, H. Villars
Identifying useful information, or tasks that inform each other	Work on discourse relation sets: ways of merging sets, of checking consistency of existing sets	Adapting to new languages (with fewer data and resources) Adapting from monologues to dialogues	Classification: cognitive impairment detection (Alzheimer, Schizophrenia) Structured prediction: argument mining
Pipeline system, Embeddings, Multi-task learning (MTL)	Clustering, few-shot learning	Cross-lingual embeddings, automatic annotations, representation learning (RL), MTL	Pipeline system, MTL, RL

+ Research Engineer (12m), PostDoc (24m), Master students (18m)

Détails du projet

- on a besoin d'informations très diverses : lexicale, morphologique, syntaxique, temporelle, sémantique...
- on ne sait pas exactement ce qu'il faut, ni ce dont manque les modèles existants
- les embeddings : très riches, très utiles, mais incomplets et comment savoir ce qui manque encore ?

→ il faut comprendre précisément les infos nécessaires / fournies. On veut tester :

- systèmes 'classiques / feature based'
- apprentissage multi-tâche : en incorporant différentes tâches liées (temps, modalité, syntaxe...) et en apprenant conjointement des tâches discursives (segmentation, attachement, relations)
- apprentissage de représentation : comparer différents embeddings, les enrichir, les fine-tuner sur des tâches liées

WP1: data representation
M. Coavoux, P. Muller
Identifying useful information, or tasks that inform each other
Pipeline system, Embeddings, Multi-task learning (MTL)

Détails du projet

W2: Frameworks divergences

N. Asher, P. Muller

Work on
discourse
relation sets:
ways of merging
sets, of checking
consistency of
existing sets

Clustering,
few-shot learning

→ pourquoi des jeux de relations différents entre les frameworks et corpus ? Peut-on aboutir à un ensemble unique ?

- étudier les divergences PDTB / RST DT
- représentation unifiée via la décomposition des relations en traits atomiques
- découverte de jeux de relations :
 - méthodes non-supervisées pour étudier comment les données se regroupent ;
 - few-shot learning basé sur une représentation unifiée des relations

→ Est-ce que les relations mal prédites correspondent à de nouvelles classes ? des erreurs d'annotation ?

→ Permettre un meilleur apprentissage joint cross-framework

Détails du projet

W3: Languages and dialogue

M. Coavoux, B. Plank

Adapting to new languages (with fewer data and resources)
Adapting from monologues to dialogues

Cross-lingual embeddings, automatic annotations, representation learning (RL), MTL

→ Construire des systèmes non dépendants ou adaptables à une nouvelle langue / domaine / modalité + profiter de toutes les données disponibles !

- nécessité d'avoir réfléchi aux questions de représentation, et ce qui est faisable sur l'anglais doit être adapté pour d'autres langues
- nécessité d'avoir réfléchi aux jeux de relations
- on veut tester :
 - embeddings cross-lingue avec fine-tuning adapté
 - RL : utiliser les embeddings 'discursifs' appris pour l'anglais ?
 - MTL : apprendre conjointement

Détails du projet

WP4: Extrinsic evaluation

M. Amblard, M. Musiol,
V. Gardette, H. Villars

Classification:
cognitive impairment
detection (Alzheimer,
Schizophrenia)
Structured
prediction: argument
mining

Pipeline system, MTL,
RL

→ L'information discursive est-elle utile ? Sous quelle forme ?

- a-t-on besoin d'un arbre complet ? de certaines relations ?
- comment incorporer l'information discursive ?
- Test de 2 tâches : analyse du langage dans le cas de troubles cognitifs et analyse argumentative

On va tester :

- pipeline avec entrée de l'information discursive sous différentes formes, et à grain plus ou moins fin
- MTL discours + application
- RL : les représentations (des mots / EDUs / phrases) apprises via une tâche discursive sont données en entrée

AnDiAMO

collaborateurs

Partner	Name	First name	Current position	Role & responsibilities
University of Toulouse - IRIT	BRAUD	Chloé	CR CNRS	Project Leader
Université de Lorraine - LORIA	AMBLARD	Maxime	MCF	Expert on dialogue and discourse modeling; semantics; Schizophrenia
University of Toulouse - IRIT	ASHER	Nicholas	DR CNRS	Expert on formal semantics and pragmatics; discourse and dialogue processing
Université Grenoble Alpes - LIG	COAVOUX	Maximin	CR CNRS	Expert on machine learning for NLP; syntactic parsing; representation learning
University of Toulouse - INSERM - CHU Toulouse	GARDETTE	Virginie	PU-PH	Expert on public health; epidemiology; Alzheimer and associated diseases
University of Toulouse - IRIT	MULLER	Philippe	MCF	Expert on machine learning for NLP; discourse processing; representation learning
Université de Lorraine - ATILF	MUSIOL	Michel	PU	Expert on cognitive psychopathology; pragmatics; Schizophrenia.
IT University of Copenhagen	PLANK	Barbara	Associate Professor	Expert on transfer learning; learning with limited data
INSERM - CHU Toulouse	VILLARS	Hélène	PH	Expert on geriatrics and gerontology; Alzheimer and associated diseases
University of Toulouse	-	-	-	Post-doc
University of Toulouse	-	-	-	Engineer

AnDiAMO

collaborateurs

Partner	Name	First name	Current position	Role & responsibilities
University of Toulouse - IRIT	BRAUD	Chloé	CR CNRS	Project Leader
Université de Lorraine - LORIA	AMBLARD	Maxime	MCF	Expert on dialogue and discourse modeling; semantics; Schizophrenia
University of Toulouse - IRIT	ASHER	Nicholas	DR CNRS	Expert on formal semantics and pragmatics; discourse and dialogue processing
Université Grenoble Alpes - LIG	COAVOUX	Maximin	CR CNRS	Expert on machine learning for NLP; syntactic parsing; representation learning
University of Toulouse - INSERM - CHU Toulouse	GARDETTE	Virginie	PU-PH	Expert on public health; epidemiology; Alzheimer and associated diseases
University of Toulouse - IRIT	MULLER	Philippe	MCF	Expert on machine learning for NLP; discourse processing; representation learning
Université de Lorraine - ATILF	MUSIOL	Michel	PU	Expert on cognitive psychopathology; pragmatics; Schizophrenia.
IT University of Copenhagen	PLANK	Barbara	Associate Professor	Expert on transfer learning; learning with limited data
INSERM - CHU Toulouse	VILLARS	Hélène	PH	Expert on geriatrics and gerontology; Alzheimer and associated diseases
University of Toulouse	-	-	-	Post-doc
University of Toulouse	-	-	-	Engineer

Discourse and dialogue

Machine / transfer learning

schizophrenia, alzheimer

Segmentation discursive sans oracle [Braud et al 2017 a,b]

- Résultats dans **4 langues et 3 domaines (anglais)**
 - En, Pt, Es, De
 - reviews (SFU), WSJ (news), instruction manuals
- Segmentation au **niveau du document**
- Utilisation d'info syntaxiques UD : POS and super tags (vs PTB info)

	Words+UD	Words+UD+STAGS
En (news)	89.0	87.0
En-SFU	87.6	86.0
En-Instr	88.3	86.4
Pt	82.9	83.0
Es	78.7	78.3
De	85.8	86.2

Architecture: bi-LSTM

- scores < SOA (sans arbres gold), bas pour petits corpus
- syntaxe pas très utile, similaire à POS

DISRPT shared task : segmenteur de discours (ToNy / DisCut)

DisRPT shared task:

- 15 corpus
- 10 langues
- styles divers (dont conversations)
- 3 formalismes: RST, PDTB, SDRT
- tout sur github (ou presque)
- format BIO unique
- **deux versions:**
 - juste tokenisé
 - ou phrase + analyse syntaxique automatique

President		BeginSeg=Yes
Bush	—	
insists	—	
it		BeginSeg=Yes
would	—	
be	—	
a		
great	—	
tool	—	
for		BeginSeg=Yes
curbing	—	
the		
budget	—	
deficit	—	
and		BeginSeg=Yes
slicing	—	
(...)		

corpus	lang	framework	train_toks	train_sents	train_docs
deu.rst.pcc	deu	rst	26,831	1,773	142
eng.pdtb.pdtb	eng	pdtb	1,061,222	44,563	1,992
eng.rst.gum	eng	rst	67,098	3,600	78
eng.rst.rstdt	eng	rst	166,849	6,672	309
eng.sdrt.stac	eng	sdrt	36,445	7,689	29
eus.rst.ert	eus	rst	21,122	990	84
fra.sdrt.annodis	fra	sdrt	22,278	880	64
nld.rst.nldt	nld	rst	17,566	1,202	56
por.rst.cstn	por	rst	44,808	1,595	110
rus.rst.rrt	rus	rst	214,484	9,859	140
spa.rst.rststb	spa	rst	43,034	1,577	203
spa.rst.sctb	spa	rst	10,249	304	32
tur.pdtb.tdb	tur	pdtb	398,203	25,080	159
zho.pdtb.cdtb	zho	pdtb	52,061	2,049	125
zho.rst.sctb	zho	rst	8,960	344	32

DISRPT shared task : segmenteur de discours (ToNy / DisCut)

DisRPT shared task:

- 15 corpus
- 10 langues
- styles divers (dont conversations)
- 3 formalismes: RST, PDTB, SDRT
- tout sur github (ou presque)
- format BIO unique
- **deux versions:**
 - juste tokenisé
 - ou phrase + analyse syntaxique automatique

Peut-on battre l'état de l'art avec des embeddings contextuels ?
Comparaison entre corpus ?

President		BeginSeg=Yes
Bush	—	
insists	—	
it		BeginSeg=Yes
would	—	
be	—	
a		
great	—	
tool	—	
for		BeginSeg=Yes
curbing	—	
the		
budget	—	
deficit	—	
and		BeginSeg=Yes
slicing	—	
(...)		

corpus	lang	framework	train_toks	train_sents	train_docs
			26,831	1,773	142
			1,061,222	44,563	1,992
			67,098	3,600	78
eng.rst.rstdt	eng	rst	166,849	6,672	309
eng.sdrt.stac	eng	sdrt	36,445	7,689	29
eus.rst.ert	eus	rst	21,122	990	84
fra.sdrt.annodis	fra	sdrt	22,278	880	64
nld.rst.nldt	nld	rst	17,566	1,202	56
por.rst.cstn	por	rst	44,808	1,595	110
rus.rst.rrt	rus	rst	214,484	9,859	140
spa.rst.rststb	spa	rst	43,034	1,577	203
spa.rst.sctb	spa	rst	10,249	304	32
tur.pdtb.tdb	tur	pdtb	398,203	25,080	159
zho.pdtb.cdtb	zho	pdtb	52,061	2,049	125
zho.rst.sctb	zho	rst	8,960	344	32

DisCut : segmentation, modèle supervisé “simple”

- information en entrée = **embedding du mot**
 - fournit robustesse sur la graphie
 - selon modèle par mot : **influence contextuelle** de la phrase +/-
Random / Glove / BeRT Multi-lingue / BeRT Mono-lingue / ELMo
- modèle séquentiel archi-classique: LSTM bidirectionnel
 - apprend influence spécifique du contexte sur la tâche
 - vers l'avant ou vers l'arrière
- sortie directe sur les labels, pas de modèle de dépendance entre labels
- Limite en nombre de tokens (from BERT):
 - pré-segmenter en phrases (CoreNLP, Stanza)
 - encode des problèmes sur certains corpus (Russian, Turkish)
- En pratique: recyclage d'un modèle de NER de AllenNLP

Comparaison sur les corpus anglais

[Xuan and Bach 2012] avec arbres syntaxiques = 93.7% → embeddings contextuels très efficaces

	Rand 50d	GloVe 50d	BERT-E	BERT-M	ELMo
PDTB	77.08	65.17	90.83	89.89	88.40
GUM	80.58	78.28	86.29	87.27	87.65
RST DT	78.97	83.21	94.41	93.72	94.75
STAC	77.43	71.70	84.65	84.45	86.06

- BERT multilingue quasi au niveau du monolingue sur l'anglais

Meilleur système en 2019 : <https://sites.google.com/view/disrpt2019/shared-task?authuser=0>

Résultats 2019

Dataset (plain)	ToNy			GumDrop			DFKI RF			Mean
	P	R	F	P	R	F	P	R	F	
deu.rst.pcc	94.88	94.49	94.68	91.99	89.80	90.88	94.20	71.77	81.47	89.35
eng.rst.gum	92.28	82.89	87.33	94.03	77.22	84.80	90.29	64.17	75.02	83.11
eng.rst.rstdt	93.60	93.27	93.43	89.56	91.43	90.49	45.96	35.85	40.28	74.87
eng.sdrst.stac	87.56	80.78	83.99	84.24	77.45	80.70	80.21	50.30	61.82	76.34
eus.rst.ert	87.43	80.94	84.06	90.06	73.36	80.86	88.21	58.01	69.99	79.21
fra.sdrst.annodis	94.31	89.15	91.65	94.46	85.29	89.64	93.47	67.35	78.29	87.07
nld.rst.nldt	94.81	89.97	92.32	94.72	88.41	91.45	95.14	68.12	79.39	88.26
por.rst.cstn	93.04	90.72	91.86	92.95	85.08	88.84	90.82	67.17	77.22	86.41
rus.rst.rrt	83.37	78.44	80.83	82.06	74.84	78.28	57.27	42.11	48.53	69.53
spa.rst.rststb	89.11	90.09	89.60	87.50	79.82	83.49	89.23	63.60	74.26	82.97
spa.rst.sctb	87.16	76.79	81.65	85.27	65.48	74.07	88.35	54.17	67.16	75.57
zho.rst.sctb	66.26	64.29	65.26	76.97	69.64	73.13	85.71	57.14	68.57	69.66
mean	88.65	84.31	86.38	88.65	79.82	83.89	83.24	58.31	68.5	80.19

Table 3: EDU segmentation results on plain tokenized data.

Transfert cross-domaine / cross-corpus

Comparaison de transfert entre corpus même langue et même formalisme

- RST DT = articles de journaux
- GUM = genres mélangés (news, académique, opinion, voyage, interviews, bio, fiction)

→ Mauvais scores : en fait des principes d'annotation assez différents (les données GUM ont d'ailleurs été ré-annotées suite à la parution de ces résultats)

Train / Test	RST DT	GUM
RST DT	93	73
GUM	66	96

Décomposition des relations de discours [Roze et al. 2019]

- plusieurs cadres théoriques avec des représentations différentes
- plusieurs schémas d'annotation / corpus
- notamment : pas de consensus sur l'ensemble de relations
 - différents niveaux de granularité, e.g. : contrast (SDRT) = Antithesis, Concession, Contrast (RST)
- pourtant encodage des mêmes informations sémantiques et pragmatiques

→ **avoir une représentation qui rend explicite cette info commune ?**

- résultats bas pour les implicites malgré une grande variété d'approches

→ est-ce que le problème vient plutôt de la représentation des données ou de

la **façon dont on modélise la tâche ?**

Modéliser la tâche différemment

- Séparer la tâche en plusieurs tâches plus simples
 - décomposer le problème
 - comprendre où sont les difficultés pour la tâche
- Décomposer l'information encodée par les étiquettes de relation en valeurs d'un petit ensemble de caractéristiques : les **primitives**

Approche :

- décomposition *a priori* des relations en primitives conceptuelles résultant de l'analyse (psycho-linguistique) [Sanders et al. 2018]
- mise à l'épreuve d'une théorie / apport théorique à un modèle empirique
- quelles primitives sont dures à prédire ?
- peut-on mieux prédire les relations ?

Exemple 1

Comparison.Concession.Contra-expectation

- (a) The biofuel is more expensive to produce.
- (b) but by reducing the tax the government makes it possible to sell the fuel the same price.
- implication attendue : 'the biofuel costs more' (Q)
- présentation d'un déni de l'attente (not-Q)
 - basic operation = causal (implique une implication, sinon additive)
 - polarity = negative (implique une négation, sinon positive)
 - implication order : seulement pour relations causales (ordre prémisse / conclusion)

Relation	Basic op	Polarity	Impl. Order	SoC	Temp
Contra-expectation	causal	neg	basic	NS	NS

Exemple 2

Source of coherence : distinction commune

- objective : contenu propositionnel
- subjective : niveau de l'acte de parole

Contingency.Pragmatic cause.Justification

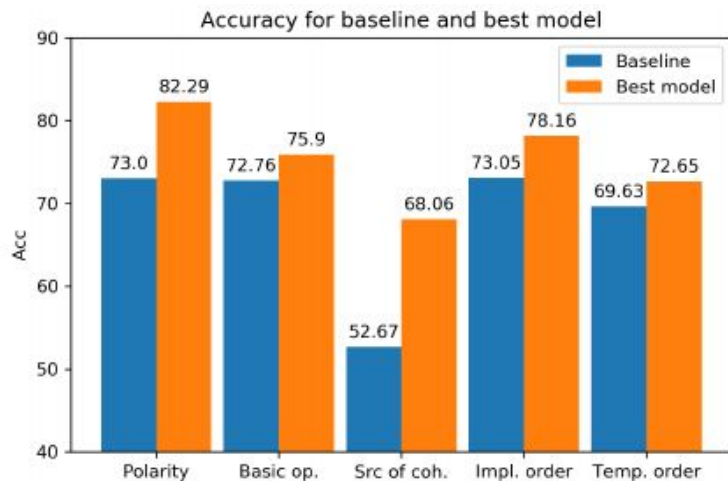
- (a) Mrs Yeargin is lying.
- (b) (because) They found students in an advanced class a year earlier who said she gave them similar help.

Relation	Basic op.	Polarity	Impl. Order	SoC	Temp
Justification	causal	pos	non basic	subjective	NS

Résultats

- basic operation
 - amélioration la + faible
 - seulement 17% des relations causales sont classées correctement → scores bas sur Contingency
- meilleur résultats pour polarity
 - 50% des relations négatives sont correctement classées
- source of coherence :
 - amélioration importante
 - mais moins d'1% de relations subjectives
- sur les relations :
 - erreur complète alors qu'1 seule primitive est fausse
 - trop de sous-spécification

→ futur : cross-corpus, MTL



	Acc	h-R	h-P	max-h-R	max-h-P
All					
Baseline	20.03	27.65	29.97	28.97	30.98
Primitives	34.15	28.89	19.32	49.07	59.05
Relations	45.35	52.97	54.95	55.42	56.58
Explicit					
Baseline	23.5	25.35	26.13	27.02	27.33
Primitives	46.27	35.56	26.43	59.93	69.59
Relations	59.08	63.63	65.3	67.4	67.8
Implicit					
Baseline	15.73	30.5	34.72	31.38	35.5
Primitives	19.12	20.63	10.52	35.61	45.99
Relations	28.35	39.76	42.11	40.57	42.67

Références (+ pub Melodi)

- Données de segmentation <https://github.com/disrpt/sharedtask2023>
- Segmenteur “utilisable” pour le Français (nouvelle version en cours de développement): <https://gitlab.irit.fr/melodi/andiamo/discoursesegmentation/discut>

Le discours est une thématique importante dans l'équipe Melodi:

- Sileo D, Van De Cruys T, Pradel C and Muller P, *Mining Discourse Markers for Unsupervised Sentence Representation Learning*, (NAACL 2019)
- Morey, M., Muller, P., and Asher, N. *A dependency perspective on RST discourse parsing and evaluation*. Computational Linguistics (2018).
- Mathieu Morey, Philippe Muller, Nicholas Asher. *How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT* (EMNLP 2018, short paper)
- Sonia Badene, Kate Thompson, Jean-Pierre Lorré and Nicholas Asher, *Weak Supervision for Learning Discourse Structure*. (EMNLP 2019)

Identification des relations (implicites)

- Implicites : considérée comme une tâche très difficile car :
 - complexe : informations à plein de niveaux (lexique, syntaxe, sémantique, connaissances du monde) + interactions entre les paires de segments / avec le reste du document
 - mais 'peu' de données (environ 18 000 annotations dans le PDTB)
 - surtout si on considère toutes les relations (distributions très déséquilibrée) → mais en fait scores très bas avec 4 classes ...
- Plein de propositions, notamment avec apprentissage par transfert :
 - [Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification](#) (2014)
 - [Comparing Word Representations for Implicit Discourse Relation Classification](#) (2015)
 - [Which aspects of discourse relations are hard to learn? Primitive decomposition for discourse relation classification](#) (2019)

Décomposition des relations de discours [Roze et al. 2019]

- plusieurs cadres théoriques avec des représentations différentes
- plusieurs schémas d'annotation / corpus
- notamment : pas de consensus sur l'ensemble de relations
 - différents niveaux de granularité, e.g. : contrast (SDRT) = Antithesis, Concession, Contrast (RST)
- pourtant encodage des mêmes informations sémantiques et pragmatiques

→ **avoir une représentation qui rend explicite cette info commune ?**

- résultats bas pour les implicites malgré une grande variété d'approches

→ est-ce que le problème vient plutôt de la représentation des données ou de

la **façon dont on modélise la tâche ?**

Modéliser la tâche différemment

- Séparer la tâche en plusieurs tâches plus simples
 - décomposer le problème
 - comprendre où sont les difficultés pour la tâche
- Décomposer l'information encodée par les étiquettes de relation en valeurs d'un petit ensemble de caractéristiques : les **primitives**

Approche :

- décomposition *a priori* des relations en primitives conceptuelles résultant de l'analyse (psycho-linguistique) [Sanders et al. 2018]
- mise à l'épreuve d'une théorie / apport théorique à un modèle empirique
- quelles primitives sont dures à prédire ?
- peut-on mieux prédire les relations ?

Cognitive Approach to Coherence Relations (CCR)

- inventaire de primitives motivées au niveau cognitif [Sanders et al. 2018]
- mappings pour PDTB, RST, SDRT
 - primitives principales [Sanders et al. 1992, 1993]
 - primitives supplémentaires : servent à expliciter les spécificités de certains cadres
- sert d'interface entre les différents cadres théoriques

Approche

- un mapping opérationnel
 - des relations annotées vers un ensemble de valeurs de primitives
 - testé sur le PDTB
- Quelles primitives sont dures à prédire ?
 - tâche de classification pour chaque primitive
- un mapping dans l'autre sens
 - d'un ensemble de valeurs de primitives vers des étiquettes de relations compatibles
 - classification de relation

Difficultés

- PDTB : hiérarchie de relation à 3 niveaux
 - 'end-labels' : + spécifique, niveau 2 ou 3
 - 'intermediate labels' : relations sous-spécifiées
- Mapping :
 - 5 primitives principales :
 - polarity, basic operation, source of coherence, implication order (2 valeurs)
 - temporal order : chronological, anti-chronological, synchronous
 - + Non-Specified : ambiguïtés (plusieurs valeurs possibles ou labels intermédiaires)
 - 3 primitives supplémentaires : conditional, alternative, specificity (binaire)
- Ces primitives ne sont pas de même importance dans le PDTB :
 - basic operation et polarity : distinctions entre classes de niveau 1
 - alternative, specificity : caractérise un ensemble plus restreint de relations
 - source of coherence : distinctions de niveau 3

Exemple 1

Comparison.Concession.Contra-expectation

- (a) The biofuel is more expensive to produce.
- (b) but by reducing the tax the government makes it possible to sell the fuel the same price.
- implication attendue : 'the biofuel costs more' (Q)
- présentation d'un déni de l'attente (not-Q)
 - basic operation = causal (implique une implication, sinon additive)
 - polarity = negative (implique une négation, sinon positive)
 - implication order : seulement pour relations causales (ordre prémisse / conclusion)

Relation	Basic op	Polarity	Impl. Order	SoC	Temp
Contra-expectation	causal	neg	basic	NS	NS

Exemple 2

Source of coherence : distinction commune

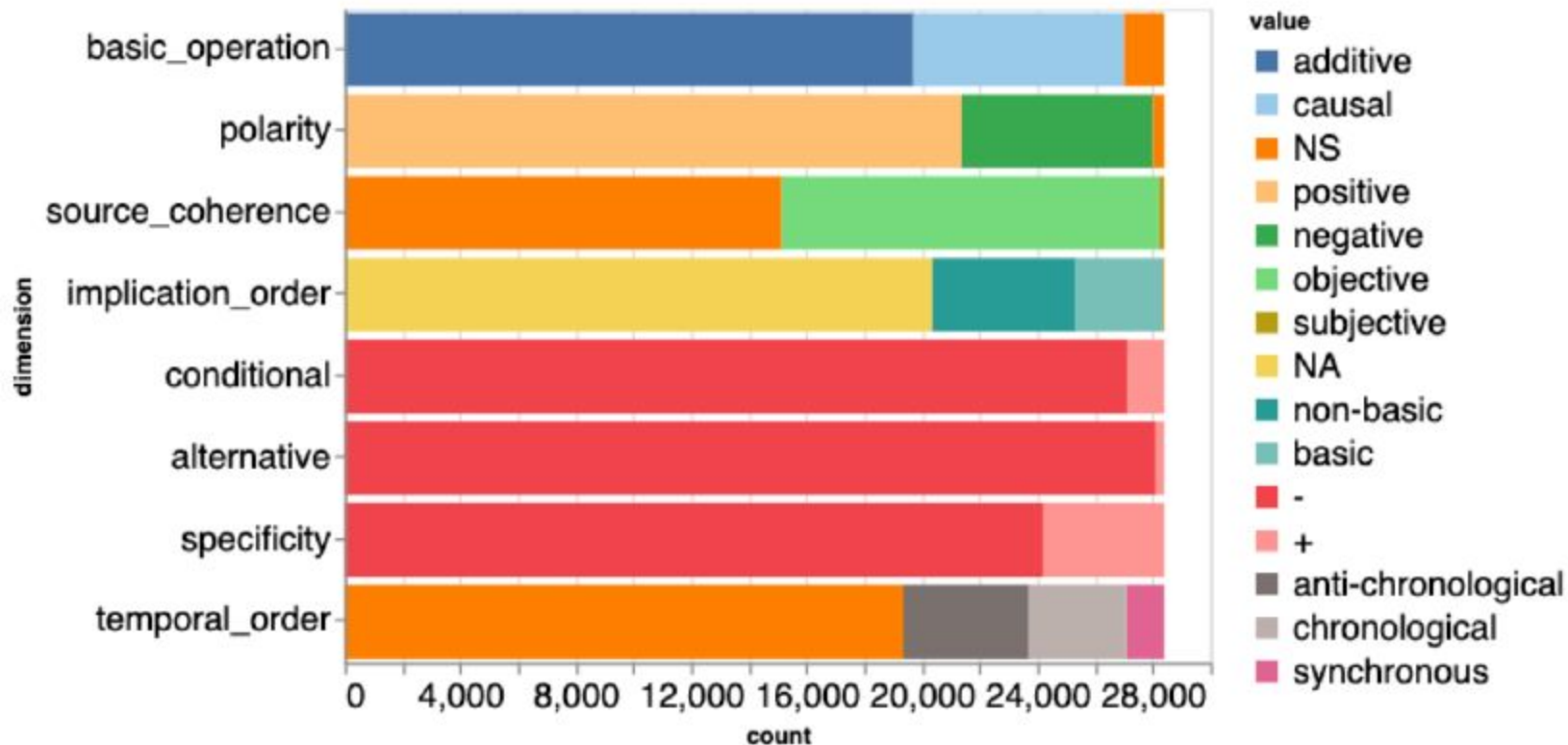
- objective : contenu propositionnel
- subjective : niveau de l'acte de parole

Contingency.Pragmatic cause.Justification

- (a) Mrs Yeargin is lying.
- (b) (because) They found students in an advanced class a year earlier who said she gave them similar help.

Relation	Basic op.	Polarity	Impl. Order	SoC	Temp
Justification	causal	pos	non basic	subjective	NS

Distribution de valeurs pour chaque primitive

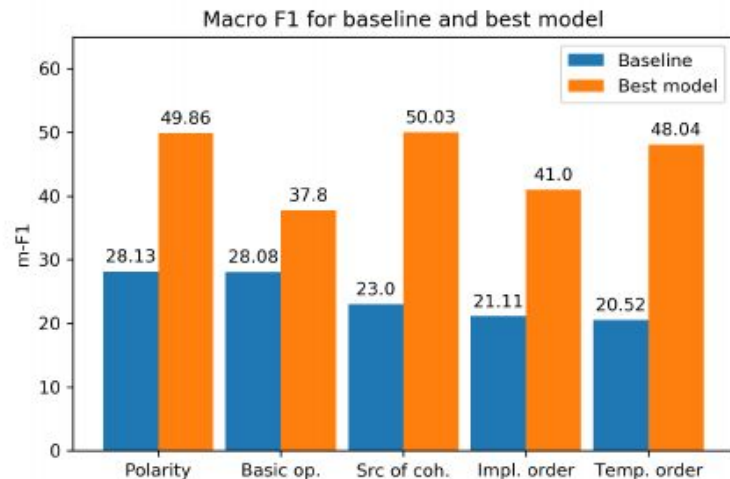
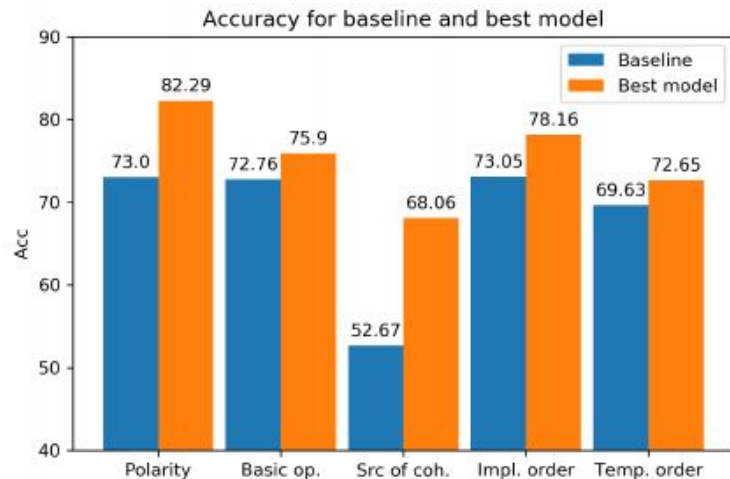


Configuration

- Classification pour chaque primitive : 28 402 paires d'arguments (train)
- Représentation de chaque argument : Infersent sentence encoder (embeddings pré-entraînés, bi-LSTM)
- Combinaison des représentations des 2 arguments :
 - concaténation
 - différence
 - produit

Résultats

- polarity et basic operation
 - primitives importantes
 - distributions similaires
- basic operation
 - amélioration la + faible
 - seulement 17% des relations causales sont classées correctement
- meilleur résultats pour polarity
 - 50% des relations négatives sont correctement classées
- source of coherence :
 - amélioration importante
 - mais moins d'1% de relations subjectives
- temporal order :
 - surtout étiquetées NS



Mapping inverse

Prédire les relations compatibles à partir des valeurs de primitives prédites

- ensemble de toutes les relations possibles (dans toute la hiérarchie)
- enlever les relations incompatibles
- enlever l'information redondante : on garde que la relation de niveau le plus haut si l'ensemble contient tous les sous-types possibles
 - Temporal, ~~Temporal.Asynchronous~~, ~~Temporal.Synchronous~~

Evaluation

Plusieurs difficultés avec l'évaluation :

- sous-spécifications (étiquette de haut niveau) : mesure de classification hiérarchique
- disjonction de relations (ensemble de relations possibles) : mesure de classification multi-label

[Kiritchenko et al. 2005]

Résultats

- trop de sous-spécification
- prédiction de trop nombreux labels
- consistant avec résultats précédents : relations de type Contingency rarement prédites
 - relations Causal pour basic operation
 - souvent Temporal prédit à la place
- erreur complète alors qu'1 seule primitive est fausse

	Acc	h-R	h-P	max-h-R	max-h-P
All					
Baseline	20.03	27.65	29.97	28.97	30.98
Primitives	34.15	28.89	19.32	49.07	59.05
Relations	45.35	52.97	54.95	55.42	56.58
Explicit					
Baseline	23.5	25.35	26.13	27.02	27.33
Primitives	46.27	35.56	26.43	59.93	69.59
Relations	59.08	63.63	65.3	67.4	67.8
Implicit					
Baseline	15.73	30.5	34.72	31.38	35.5
Primitives	19.12	20.63	10.52	35.61	45.99
Relations	28.35	39.76	42.11	40.57	42.67

Finalemment :

Primitives difficiles à prédire, qui rendent la tâche difficile ?

- basic operation : l'une des + importantes très difficile à prédire

Les primitives ne sont pas indépendantes les unes des autres

- les apprendre en isolation est forcément moins précis que d'apprendre directement les étiquettes complètes
- futur : apprentissage multi-tâche

Autre extension :

- cross-corpus (RST et SDRT)
- few-shot : apprentissage sans certaines relations, peut-on les prédire quand même ?