État de l'art sur les marqueurs discursifs en Traitement Automatique des Langues

Fatou Sow[1, 2]
(1) LORIA, Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
(2) ATILF, Université de Lorraine, CNRS, F-54000 Nancy, France
fatou.sow@loria.fr

RÉSUMÉ

Les marqueurs discursifs sont des éléments linguistiques qui peuvent être employés pour construire la cohérence d'un discours car ils expriment les relations entre les unités discursives. Ils constituent ainsi des indices utiles pour la résolution de problèmes de traitement de langue en rapport avec la sémantique du texte, le discours ou la compréhension de systèmes. Dans cet article, nous présentons un état de l'art des marqueurs discursifs en traitement automatique des langues (TAL). Nous introduisons les représentations textuelles des marqueurs discursifs puis nous nous intéressons à la détection des marqueurs et l'utilisation de leurs sens pour améliorer ou évaluer des tâches de TAL.

ABSTRACT

Literature review on discourse markers in Natural Language Processing

Discourse markers are defined as linguistic items that can be used to form a coherent discourse because they express the relationship between discourse units. They are signals that are useful for solving natural language processing (NLP) tasks related to text semantics, discourse or system understanding. This paper presents a literature review on discourse markers in natural language processing. We focus on textual representations of discourse markers and then, we investigate the detection of markers and the use of their meanings to improve or evaluate NLP tasks.

MOTS-CLÉS : marqueurs discursifs, état de l'art, multilingue.

KEYWORDS: discourse markers, state of the art, multilingual.

1 Introduction

Dans un but communicatif, les locuteurs d'une langue élaborent un discours dont le degré de cohérence va être déterminant pour la compréhension par un interlocuteur. Ils vont ainsi user de différents procédés linguistiques tels que l'utilisation de marqueurs discursifs pour structurer leur propos.

Les marqueurs du discours sont des éléments linguistiques qui ont donné naissance à de nombreuses définitions en pragmatique. En particulier, Schiffrin (1987) est la première à se pencher en profondeur sur le concept de *discourse markers* et les définit comme des « éléments séquentiellement dépendants qui délimitent des unités de parole ». À sa suite, plusieurs linguistes ont proposé des définitions différentes, ce qui a conduit à une pluralité de termes qui se chevauchent. Il est ainsi question de marqueurs discursifs, de connecteurs, de particules, de particules énonciatives, de signaux du discours, etc. (Dostie, 2004). De plus, les propriétés discursives des marqueurs changent selon la terminologie

choisie. Ainsi, certains auteurs vont associer les marqueurs discursifs à l'expression d'une relation du discours comme l'illustre l'exemple (1) avec le marqueur discursif *même si* qui représente la relation de concession. Ce type de marqueur discursif est aussi communément appelé connecteur du discours. D'autres auteurs vont souligner l'usage des marqueurs d'un point de vue conversationnel et qui relève plutôt du sentiment du locuteur. Dans ce cas, il est souvent question de particule du discours, comme indiqué en (2) avec la particule *eh bien* qui rend compte de la surprise du locuteur.

(1) *Elle est sortie se promener* ***même si*** *la pluie ne s'est pas arrêtée.*

(2) ***Eh bien*** *je ne pensais pas que tu viendrais.*

En traitement automatique des langues, les marqueurs du discours sont globalement associés à la notion de connecteurs du discours et de nombreuses recherches ont ainsi été effectuées pour identifier et exploiter leurs propriétés discursives, à travers des tâches aussi bien manuelles qu'automatiques. C'est ainsi que dans la suite de l'article, par souci d'homogénéité, nous utiliserons le terme « marqueur discursif » pour englober tous les concepts qui lui sont afférents, y compris les connecteurs du discours.

À notre connaissance, nous présentons le premier état de l'art portant sur les marqueurs discursifs dans le domaine du TAL. D'abord, nous nous intéressons aux ressources textuelles en lien avec les marqueurs discursifs. Ensuite, nous nous penchons sur les méthodes d'identification des marqueurs discursifs et leurs usages dans d'autres tâches. Enfin, nous étudions l'emploi de la sémantique des marqueurs discursifs pour améliorer des systèmes de TAL. Par ailleurs, dans ce papier, nous considérons les marqueurs discursifs dans différentes langues.

2 Les ressources lexicales associées aux marqueurs discursifs

2.1 L'annotation des marqueurs dans les corpus discursifs

Corpus	Langues	Nombre phrases	Nombre tokens	Nombre annotations
PDTB 3.0	anglais	-	-	53 631
Discovery	anglais	3 480 000	-	1 740 000
TED-MDB	anglais	-	7 012	488
	russe	-	5 623	458
	polonais	-	6 520	413
	portugais	-	7 166	525
	allemand	-	6 366	454
	turc	-	5 164	478
DISRPT	anglais	60 667	1 368 584	32 904
	italien	3 753	26 114	1 071
	portugais	5 588	195 039	5 464
	thaï	6 534	256 523	10 864
	turc	31 606	493 532	9 130
	chinois	2 891	73 314	1 660

TABLE 1 – Distribution des tokens et des annotations de marqueurs discursifs par corpus

Différentes formalisations du discours ont été appliquées à des données textuelles pour former des corpus annotés discursivement. Ces corpus peuvent ainsi comporter des annotations de marqueurs discursifs selon la méthodologie adoptée.

C'est le cas du Penn Discourse Treebank (PDTB) (Miltsakaki *et al.*, 2004) qui est un projet d'annotation qui s'appuie sur les données du Penn Treebank, un large corpus anglais syntaxiquement étiqueté. Dans sa première version, 10 marqueurs discursifs et leurs arguments ont été sélectionnés et annotés, un argument étant défini comme une expression qui comporte un prédicat. Un argument peut aussi être exprimé par une phrase nominale ou un déictique qui exprime « un événement ou un état ». De plus, le concept de marqueurs discursifs implicites est défini dans le PDTB : pour une paire d'arguments liée par une relation du discours qui n'est pas textuellement exprimée, les annotateurs rajoutent un marqueur discursif qui correspondrait le mieux à cette relation. La notion d'argument et de marqueur discursif implicite est indiquée en (3).

(3) [*The $6 billion that some 40 companies are looking to raise in the year ending March 31 compares with only $2.7 billion raised on the capital market in the previous fiscal year*]$_{arg}$. ***IMPLICIT-(In contrast)*** [*In fiscal 1984 before Mr. Gandhi came to power, only $810 million was raised*]$_{arg}$ (Miltsakaki *et al.*, 2004)

De la première version à la deuxième version, le PDTB 2.0 (Prasad *et al.*, 2008) s'enrichit d'une annotation de 100 marqueurs discursifs explicites différents et 102 marqueurs discursifs implicites, pour un total de 18 459 occurrences de marqueurs discursifs explicites et 16 053 marqueurs discursifs implicites annotés [1]. De plus, une définition hiérarchique des relations du discours est proposée.

Enfin, une troisième version a été publiée par Prasad *et al.* (2018) avec la modification de certaines relations et l'ajout de nouveaux types de relations, portant le nombre d'annotations à 53 631 tokens [2]. Le PDTB a été adapté à plusieurs langues telles que le chinois (Zhou & Xue, 2015) ou encore le hindi (Oza *et al.*, 2009). Par ailleurs, le projet du PDTB pour le français, le French Discourse Treebank, a été défini par Danlos *et al.* (2012) et l'objectif était d'annoter le French TreeBank, en s'appuyant sur le lexique de marqueurs discursifs du LEXCONN de Roze *et al.* (2012) pour identifier les marqueurs discursifs. L'idée du FDTB est alors de reprendre la méthodologie du PDTB tout en modifiant la structure des relations en s'inspirant de la RST (Mann *et al.*, 1989) et de la SDRT (Lascarides & Asher, 2007). En outre, Danlos *et al.* (2012) proposent une annotation de tous les marqueurs discursifs implicites et de tous les arguments, contrairement au PDTB. Le projet est toujours en cours, avec une première étape qui a été publiée par Danlos *et al.* (2015).

En plus du PDTB qui est un corpus annoté pour l'anglais, un autre corpus en anglais appelé Discovery a été constitué automatiquement par Sileo *et al.* (2019). À partir du depCC (Panchenko *et al.*, 2018), un large corpus issu des données du web et annoté syntaxiquement, un ensemble de paires de phrases lié par un marqueur discursif candidat est extrait. Le marqueur discursif candidat, qui peut être un adverbe ou une conjonction est au début de la deuxième phrase de la paire. Ensuite, à l'aide d'un classifieur de FastText [3], ils effectuent une prédiction du marqueur discursif à partir de la paire ou de la deuxième phrase. Les exemples comportant les marqueurs discursifs les plus prédits sont exclus du jeu de données. En effet, Sileo *et al.* (2019) considèrent que les marqueurs discursifs candidats qui sont facilement détectables par des modèles lexicaux simples, tels que FastText, constituent

1. https://catalog.ldc.upenn.edu/LDC2008T05
2. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf
3. https://fasttext.cc

probablement du bruit. Ils partent du principe que les marqueurs discursifs ont des propriétés plus complexes, rendant leur détection moins évidente. Le corpus Discovery comporte 174 marqueurs du discours différents pour 1,74 million de paires de phrases, sous la forme indiquée en (4). Il s'agit d'un corpus de grande taille comportant de nombreux exemples d'usage de marqueurs discursifs en anglais. Cependant, son annotation reste moins fine que le PDTB qui prend en compte les relations du discours et les marqueurs discursifs implicites.

(4) *The product is set to launch in about 40 countries on November 1st.* | *One of those countries will be China, which has never before been a launch region for the iPad.* | ***significantly,*** (Sileo *et al.*, 2019)

En parallèle de la constitution de grands corpus discursifs monolingues, et en anglais principalement, il existe des initiatives pour constituer des corpus multilingues afin de représenter un plus grand nombre de langues. Par exemple, le TED Multilingual Discourse Bank ou TED-MDB (Zeyrek *et al.*, 2018) consiste en 36 transcriptions de conférences TED en anglais, allemand, polonais, portugais, russe et turc. Ces transcriptions ont été annotées discursivement à l'aide du cadre théorique du PDTB et dans la mesure où le registre des conférences TED est celui de l'oral préparé, les annotations ont cherché à en rendre compte en annotant des phénomènes tels que l'hypophore. Il s'agit d'un cas de figure où le locuteur répond à sa propre question, comme l'illustre l'exemple (5).

(5) *Why is that hard ? Well to see, let's imagine we take the Hubble Space Telescope and we turn it around ... We'll see something like that, a slightly blurry picture of the Earth. ..* ***[AltLex : Hypophora]*** *[En]* (Zeyrek *et al.*, 2018)

En plus de la question du multilinguisme, une autre problématique dans l'analyse du discours est la diversité des théories. Ceci a conduit à la création de plusieurs corpus de petite taille aux types d'annotations différentes. Néanmoins, à l'ère actuelle des modèles de langues, de grandes quantités de données structurées sont nécessaires pour leurs entraînements. Le projet DISRPT va ainsi proposer un corpus de référence multilingue qui va regrouper différentes méthodologies en un format commun. Braud *et al.* (2024), à travers trois éditions, vont collecter 28 jeux de données annotés dans les cadres théoriques de la RST (Mann *et al.*, 1989), la SDRT (Lascarides & Asher, 2007), le PDTB (Miltsakaki *et al.*, 2004) et le Dependency (Li *et al.*, 2014). Ils vont ensuite homogénéiser l'annotation des unités discursives, des marqueurs discursifs et des relations du discours. DISRPT est disponible en 13 langues, à savoir l'anglais, l'allemand, le français, le portugais, l'espagnol, le russe, le chinois, l'italien, le turc, le néerlandais, le thaï, le basque et le persan. Par ailleurs, les corpus spécifiquement annotés en marqueurs discursifs sont au nombre de six, comme indiqué en 1. De plus, une compétition (*Shared Task*) est proposée avec trois tâches : la détection de marqueurs discursifs, la segmentation en unités discursives et la classification des relations du discours. Pour la tâche d'identification de marqueurs discursifs, les données sont annotées avec les étiquettes BIO qui différencient les marqueurs discursifs, annotés B (*Beginning*) ou I (*Inside*), des autres tokens étiquetés O (*Outside*). De plus, il existe deux formats de données, celles sans annotations syntaxiques et celles qui incluent des annotations syntaxiques.

Ainsi, plusieurs corpus ont été constitués pour représenter les éléments discursifs des textes. Ces corpus, disponibles dans différentes langues, sont largement utilisés dans plusieurs tâches de TAL telles que la détection de marqueurs discursifs (3), celle de relations implicites (4.2) ou l'implémentation d'analyseurs discursifs. Ils sont en outre utilisés pour la constitution de lexiques de marqueurs

discursifs (2.2). Le tableau 1 récapitule la taille des corpus présentés et le nombre d'annotations obtenues.

2.2 La constitution de lexiques de marqueurs discursifs

Différents projets ont vu le jour pour constituer des lexiques de marqueurs discursifs dans différentes langues plus ou moins dotées. Pouvant être créés manuellement ou automatiquement, ils tendent à représenter les usages spécifiques des marqueurs discursifs. Initialement pertinents pour les tâches de génération de texte, les lexiques peuvent être employés pour des études linguistiques ou fournir une base pour l'analyse de marqueurs discursifs dans une langue.

À ce titre, Stede & Umbach (1998) présentent DiMLex qui est initialement une méthodologie pour définir et extraire des marqueurs discursifs, dans le but de constituer un lexique en allemand. Pasch *et al.* (2003) proposent ensuite un ensemble de critères pour identifier un marqueur du discours. Il s'agirait d'une expression lexicale invariable qui représente une relation à deux arguments, les arguments de la relation étant des propositions. À partir de ces critères et de ressources textuelles en allemand, une extraction manuelle des marqueurs discursifs est effectuée et les marqueurs discursifs obtenus sont annotés avec différents attributs tels que la catégorie syntaxique, les différentes orthographes possibles du marqueur, les relations du discours qui lui sont associées et des exemples d'usage.

La définition proposée par DiMLex ainsi que son format de représentation des attributs des marqueurs ont été repris dans plusieurs langues. Das *et al.* (2018) présentent ainsi une version de DiMLex en anglais et énumèrent 149 marqueurs discursifs en s'appuyant sur sa méthodologie et le PDTB. La version turque de DiMLex va quant à elle modifier certaines règles de DiMLex pour répondre aux réalités de la langue : Zeyrek & Başıbüyük (2019) vont par exemple prendre en compte les marqueurs discursifs suffixaux qui, contrairement aux marqueurs discursifs en allemand, changent de forme. Un lexique de 113 marqueurs discursifs a ainsi été constitué avec des exemples extraits des corpus TDB (Zeyrek & Kurfalı, 2017) et TED-MDB. Feltracco *et al.* (2016) vont aussi adopter la méthodologie de DiMLex en adaptant les critères à l'italien, de même que pour le lexique CzeDlex (Mírovský *et al.*, 2017) qui par ailleurs utilise le Prague Discourse Treebank 2.0 (Rysová *et al.*, 2016) pour compléter les annotations des marqueurs discursifs. Mendes *et al.* (2018) vont également prendre parti des annotations du TED-MDB pour compléter un lexique de marqueurs discursifs en portugais européen. Il est ainsi important de noter que les corpus discursifs tels que le PDTB en anglais ou le TED-MDB facilitent la collecte préalable d'une liste de marqueurs candidats ainsi que des exemples d'usage discursif des marqueurs. En outre, l'existence de ces corpus facilite la possibilité d'avoir une approche manuelle dans la constitution des lexiques.

Dans le cas des langues qui ont peu de ressources discursives telles que le bengali, Das *et al.* (2020) vont s'appuyer sur les traductions de la version anglaise du DiMLex pour extraire une liste de marqueurs candidats du corpus *Bangla RST DT* (Das & Stede, 2018). 123 marqueurs discursifs sont finalement obtenus. Pour le pidgin nigérien qui est aussi une langue peu dotée et ne comporte pas de corpus annoté discursivement, une approche semi-automatique a été nécessaire pour construire un lexique. Marchal *et al.* (2021) vont ainsi utiliser le Naija Treebank (Caron *et al.*, 2019), un corpus parallèle du pidgin nigérien avec une traduction en anglais. Ils vont ensuite appliquer un analyseur discursif sur la traduction anglaise pour identifier les marqueurs discursifs et extraire manuellement les marqueurs discursifs candidats correspondants en pidgin nigérien. Un ensemble de filtres syntaxiques et statistiques ont été appliqués, résultant en une liste de 57 marqueurs discursifs. Bourgonje *et al.* (2018) emploient également un corpus parallèle ainsi que la version allemande de DiMLex pour

recenser les marqueurs discursifs en néerlandais et les représenter au format DiMLex.

Pour le français, LEXCONN constitue le lexique de marqueurs discursifs de référence, qui s'appuie sur les données du corpus Frantext et d'un ensemble de critères sémantiques, syntaxiques et discursifs définis par Roze *et al.* (2012). Il comporte 328 marqueurs discursifs annotés avec leurs catégories syntaxiques et les relations du discours associées. Par ailleurs, LEXCONN a été utilisé avec DiMLex pour construire un lexique bilingue (Rauh *et al.*, 2023).

2.3 L'accessibilité des ressources

Les ressources citées précédemment ont été implémentées pour enrichir les informations sur les marqueurs discursifs dans différentes langues. Ces données sont disponibles numériquement et des projets ont ainsi cherché à les rassembler sur une seule plateforme, pour faciliter leur accès. Ainsi, la plateforme Connective-Lex [4] recense les différentes versions du DiMLex ainsi que d'autres lexiques tels que le LEXCONN et le Naija-Lex de Marchal *et al.* (2021). Les lexiques peuvent ainsi être consultés avec des options de filtre sur la catégorie syntaxique des marqueurs ou la relation qu'ils expriment.

En dehors de la plateforme en ligne, un autre format possible est la représentation des marqueurs discursifs sous forme de données du Web Sémantique. Ainsi, Chiarcos & Ionov (2021) ont proposé d'encoder des lexiques multilingues de marqueurs du discours, tels que DiMLex, LexCONN ainsi que le PDTB, au format RDF. Ce format, interprétable par la machine, facilite ainsi le requêtage sur les lexiques tout en proposant une formalisation conçue pour s'adapter à de nouveaux lexiques. Par ailleurs, Chiarcos (2022) prend parti de cette ressource et, à l'aide de dictionnaires bilingues au format RDF, il construit un lexique de marqueurs du discours pour neuf langues (bulgare, grec, esperanto, finnois, japonais, norvégien, polonais, suédois, turc) par induction.

3 Les méthodes d'identification des marqueurs discursifs

Dans le domaine du TAL, les marqueurs discursifs sont principalement étudiés à travers la tâche de prédiction qui peut prendre deux formes. Une possibilité est d'identifier, par une classification multi-classes, le marqueur discursif entre une paire de phrases ; chaque classe à identifier est un marqueur du discours. C'est le cas de Malmi *et al.* (2018), qui vont effectuer une classification de 20 classes avec 19 marqueurs discursifs et une classe pour représenter l'absence de marqueurs. Les données consistent en paires de phrases de Wikipédia en anglais qui comportent ou non un marqueur discursif entre les deux phrases. À partir de 400 000 paires de phrases, ils entraînent un modèle de décomposition d'attention dont l'architecture permet d'obtenir des plongements lexicaux qui représentent les relations entre les tokens de la paire de phrases et celles des tokens par rapport à l'autre phrase de la paire. La décomposition d'attention s'appuie sur le calcul de scores d'attention et des vecteurs de comparaison et avec cette méthode, Malmi *et al.* (2018) obtiennent une exactitude de 32,71 pour une F-mesure de 31,80. Par ailleurs, afin de positionner leur performance par rapport aux humains, ils ont constitué un jeu de données de 10 000 paires de phrases évaluées par trois annotateurs qui devaient choisir le marqueur discursif adéquat pour chaque paire de phrase. Avec une exactitude de 23,12 et une F-mesure de 23,72, ces résultats indiqueraient que les humains ont relativement des

4. http://connective-lex.info

difficultés à détecter le marqueur discursif adéquat et que leur identification serait une tâche complexe pour les locuteurs. Cependant, ce jeu de données annoté manuellement est possiblement soumis à des biais dans la mesure où les rédacteurs des articles de Wikipédia, dans un souci de clarté, auraient tendance à avoir une utilisation supérieure à la normale des marqueurs discursifs.

La deuxième manière de prédire les marqueurs discursifs est d'identifier les occurrences de marqueurs dans une séquence de tokens, par exemple à l'aide de l'étiquetage BIO. Ainsi, le projet DISRPT présenté en 2.1 adopte ce format pour les données et la constitution de corpus va de pair avec une compétition dont les trois éditions comportent une tâche de détection de marqueurs discursifs. Pour la première édition (Zeldes *et al.*, 2019), les architectures utilisées étaient essentiellement des réseaux de neurones récurrents (RNN) ou des arbres de décision. Le meilleur modèle s'appuyait sur un RNN avec des plongements contextualisés de BERT et obtient une moyenne F-mesure de 79,25 (données avec annotations syntaxiques) et 83,63 (données sans annotations syntaxiques) pour l'anglais, le turc et le chinois. De même, dans les éditions suivantes de Zeldes *et al.* (2021) et Braud *et al.* (2023), les modèles les plus performants utilisent essentiellement les plongements contextualisés de BERT. C'est le cas de HITS, le modèle le plus performant de 2023 (Liu *et al.*, 2023) qui associe les plongements lexicaux de BERT à une couche bi-LSTM et une couche CRF adaptée aux tâches d'étiquetage de tokens. HITS atteint une moyenne F-mesure de 80,47 pour les données avec des annotations syntaxiques pour l'anglais, l'italien, le portugais, le thaï, le turc et le chinois.

Chapados Muermans & Kosseim (2022) utilisent aussi les données de DISRPT pour implémenter un modèle multilingue pour l'anglais, le turc et le chinois. Il s'agit d'un affinage de BERT avec les données de DISRPT, auquel une couche CRF est ajoutée. Par ailleurs, ils créent des données synthétiques pour évaluer si l'ajout de données augmente les performances de BERT. À partir de corpus alignés en anglais-turc et en anglais-chinois où leur modèle de détection de marqueurs discursifs a été appliqué, deux types de jeux de données ont été constitués. D'une part, les données de projection sont constituées des annotations issues du modèle en turc/chinois et des marqueurs discursifs uniquement annotés en anglais qui sont projetés sur les mots correspondants en turc ou chinois. D'autre part, les données d'alignement sont les séquences dont les marqueurs discursifs ont été annotés aussi bien en anglais qu'en turc ou en chinois. Le modèle de détection de marqueurs discursifs obtient cependant de meilleures performances, sans ajout des données synthétiques, avec en anglais une F-mesure de 92,49 (±0,77), en turc une F-mesure de 93,97 (± 0,34) et en chinois une F-mesure de 87,47 (±0,95). La chute de performance par rapport à un entraînement avec uniquement les données de DISRPT laisse penser que les données rajoutées créent du bruit et que l'affinage de BERT ne nécessite pas l'usage d'une grande quantité de données. Il faut cependant noter que les données d'évaluation et celles d'entraînement sont issues du même corpus, contrairement aux données d'alignement, ce qui pourrait également expliquer les résultats moins élevés.

Par ailleurs, dans l'état de l'art, les marqueurs discursifs discontinus, tels que *d'une part ... d'autre part* en français, sont rarement pris en compte dans la détection de marqueurs car ils soulèvent des problèmes lors de la tokenisation ou de la représentation des données en plongements lexicaux. Par exemple, Braud *et al.* (2024) annotent les marqueurs discursifs discontinus en deux marqueurs différents. Costa *et al.* (2023) étudient ce problème en effectuant une tâche de détection de ces marqueurs discursifs en chinois. D'abord, ils ont effectué une annotation manuelle au format BIO des marqueurs discontinus dans le CDTB. Ensuite, ils ont appliqué d'une part des algorithmes d'apprentissage automatique (arbre de décision, SVM, Perceptron, forêt aléatoire) et d'autre part l'*hypothese testing* en s'appuyant sur les collocations. Les meilleures performances ont été obtenues avec SVM et la forêt aléatoire avec une F-mesure de 75,6. En comparaison aux autres modèles s'appuyant sur BERT, Costa *et al.* (2023) obtiennent des performances moindres. Il serait donc

intéressant de l'appliquer pour la tâche de prédiction de marqueurs discontinus, ce qui de plus permettrait d'évaluer la difficulté de la détection de ce type de marqueurs par rapport aux marqueurs consécutifs.

Les résultats de ces différents modèles de détection de marqueurs discursifs démontrent que les modèles basés sur l'architecture *Transformer* obtiennent de meilleures performances que les autres architectures neuronales. En outre, le choix d'une méthode de prédiction est dépendant des données utilisées pour l'entraînement des modèles car la nature de la tâche d'identification dépend du type d'annotation choisie.

4 Les enjeux associés à l'identification de marqueurs discursifs

La tâche de détection de marqueurs discursifs va souvent de pair avec une tâche de TAL et dans ces cas de figure, la détection de marqueurs est utilisée pour augmenter les performances d'une tâche principale. Elle peut ainsi être employée pour améliorer des plongements lexicaux ou faciliter la détection de relation implicite. En outre, la tâche de prédiction de marqueurs discursifs peut être employée pour évaluer les analyseurs discursifs ou diagnostiquer le fonctionnement d'un modèle.

4.1 Pour de meilleurs plongements lexicaux

Les plongements lexicaux sont un format vectoriel adapté aux algorithmes d'apprentissage automatique ou profond et l'objectif des modèles de plongements lexicaux est d'encoder les propriétés sémantiques d'un mot ou d'une phrase. Ainsi, pour améliorer l'entraînement de ces modèles, une possibilité est d'ajouter une tâche de prédiction de marqueur discursif car cela constitue un moyen d'intégrer dans les plongements obtenus les rapports discursifs entre les mots.

Wu *et al.* (2019) ont ainsi conçu des modèles de plongements lexicaux de mots pour la classification de relations implicites et y intègrent la détection de marqueurs discursifs comme une tâche de classification. Ils font le choix d'une approche neuronale et proposent deux modèles. Le premier modèle, dit *Average Model*, encode les arguments d'un marqueur discursif puis représente leur moyenne ; le plongement obtenu est ensuite donné en entrée à un Perceptron multi-couches pour la détection de marqueurs discursifs. Le deuxième modèle, *Interaction Model*, représente plutôt les relations entre les paires de mots des arguments avant d'encoder les arguments du marqueur discursif dans une couche d'agrégation puis d'effectuer une tâche de détection de marqueur discursif avec Perceptron. La tâche de classification de marqueurs discursifs et les architectures employées pour la représentation des mots permettent de construire des plongements lexicaux qui capturent les relations discursives entre les mots, ce qui participe à améliorer la détection de relation implicite. Wu *et al.* (2019) obtiennent une meilleure détection de marqueurs discursifs en utilisant l'*Interaction Model* avec une exactitude de 58,3% pour 96 marqueurs discursifs, 58,9% pour 60 marqueurs, 60,8% pour 30 marqueurs, 62,8% pour 20 marqueurs et 69,9% pour 10 marqueurs. L'exactitude du modèle est ainsi corrélée au nombre de marqueurs discursifs à détecter.

Selon la tâche, par exemple pour des tâches de questions-réponses ou encore la génération de résumés, il est pertinent d'avoir une représentation à l'échelle de la phrase plutôt qu'à celle du mot. C'est ainsi que Nie *et al.* (2019) proposent DisSent, deux versions d'architectures de plongements lexicaux de phrases entraînés avec la tâche de prédiction de marqueurs discursifs. En effet, ils considèrent que

la capacité du modèle à identifier les marqueurs discursifs va en parallèle enrichir sémantiquement les plongements de phrases. Pour l'architecture, ils s'inspirent du modèle InferSent de Conneau *et al.* (2017) et implémentent un modèle qui s'appuie sur un bi-LSTM et qui représente la paire d'arguments d'un marqueur discursif en un plongement lexical. Ce plongement est ensuite donné en entrée à une couche dense qui effectue la prédiction de marqueurs discursifs. Leur deuxième modèle reprend aussi la tâche de prédiction, avec au préalable un affinage de BERT qui consiste à prédire un marqueur discursif à partir d'une paire de phrases en entrée. Les données d'entraînement pour les deux architectures sont des paires de phrases issues du BookCorpus de Zhu *et al.* (2015), extraites avec un analyseur syntaxique implémenté par Nie *et al.* (2019). Les phrases d'une paire sont liées par un marqueur discursif et trois jeux de données sont constitués et prennent en compte 5, 8 et 15 marqueurs discursifs issus du PDTB. De même que Wu *et al.* (2019), Nie *et al.* (2019) obtiennent des résultats de plus en plus exacts en réduisant le nombre de marqueurs à détecter. Ainsi, DisSent parvient à une exactitude de 86,1 (F-mesure : 82,6) pour la détection de 5 marqueurs discursifs, 82,9 (F-mesure : 76,2) pour 8 marqueurs discursifs et 77,5 (F-mesure : 60,1) pour 15 marqueurs discursifs.

4.2 L'identification des marqueurs discursifs au service de la détection de relations implicites

Dans le cas où deux unités discursives sont liées par une relation du discours en l'absence de marqueurs discursifs, la relation est dite implicite. L'identification d'une relation implicite est une question de recherche toujours actuelle dans la mesure où l'absence de marqueurs discursifs complexifie leur détection par les modèles automatiques. De plus, les ressources qui prennent en compte les annotations de relations implicites, telles que le PDTB, comportent peu d'annotations pour permettre un entraînement robuste des grands modèles de langue. Pour pallier cette absence de données, une alternative est d'entraîner les modèles sur des paires de phrases explicites dont le marqueur discursif a été au préalable supprimé (Marcu & Echihabi, 2002). Cependant, Sporleder & Lascarides (2008) ont montré que les données explicites diffèrent sémantiquement des relations implicites et ainsi, l'utilisation de données explicites pour la détection de relations implicites ne permet pas une généralisation de la tâche. D'autres stratégies vont alors être implémentées pour incorporer la prédiction de marqueurs discursifs dans le modèle de détection de relations implicites. L'objectif est alors d'enrichir les compétences discursives du modèle pour améliorer ses performances.

(6) *But a few funds have taken other defensive steps. Some have raised their cash positions to record levels.* ***Implicit = BECAUSE*** *High cash positions help buffer a fund when the market falls.* (Prasad *et al.*, 2008) → Relation : Contingency.Cause

Il est important de noter que pour la tâche de détection de relations implicites, le PDTB est le corpus le plus utilisé du fait de ses annotations de marqueurs discursifs implicites. Comme illustré en (6), dans le cas où une relation du discours est exprimée entre deux arguments malgré l'absence de marqueurs discursifs, les annotateurs du PDTB insèrent un marqueur discursif qui représenterait la relation puis indiquent la relation associée à ce marqueur discursif. À partir de ces annotations, Lin *et al.* (2009) proposent deux découpages des données pour l'entraînement de modèles de détection de relations implicites. Ils considèrent que pour une évaluation de cette tâche, il est possible d'effectuer une classification des relations de premier niveau du PDTB 2.0 (*Comparison, Expansion, Contingency, Temporal*) ou d'effectuer une classification plus fine avec la détection des relations de deuxième niveau du PDTB 2.0. Cependant, pour ce type de relations, Lin *et al.* (2009) sélectionnent celles qui

ont suffisamment d'exemples pour former des données d'entraînement ; elles sont au nombre de 11 : *Asynchronous, Synchrony, Cause, Pragmatic Cause, Contrast, Concession, Conjunction, Instantiation, Restatement, Alternative, List*. Ce découpage a ainsi été largement utilisé pour la tâche de détection de relations implicites.

Une stratégie possible pour la détection de relations implicites est d'utiliser un modèle pré-entraîné pour des tâches générales de TAL et de le spécialiser pour les tâches discursives, tout en conservant ses performances générales. Ainsi, Kishimoto *et al.* (2020) vont chercher à optimiser BERT pour la détection de relations implicites grâce à la tâche de détection de marqueurs discursifs. Ils vont ainsi proposer deux configurations en lien avec cette tâche. La première intègre dans le pré-entraînement de BERT la détection de marqueur discursif explicite, c'est-à-dire qu'une paire d'arguments sans le marqueur discursif est donnée en entrée et le modèle apprend à prédire le marqueur discursif. La deuxième configuration consiste à détecter le marqueur discursif implicite de la paire d'arguments donnée en entrée ; cette tâche est un affinage de BERT. Différentes configurations ont été testées et les résultats avec la classification de relations de premier niveau du PDTB 2.0 indiquent que BERT intégrant la tâche de détection de marqueur discursif explicite donne de meilleurs résultats, avec une exactitude de 65,26 (F-mesure : 58,48). Pour les relations de deuxième niveau, BERT associé aux deux tâches donne une F-mesure de 51,79.

Une autre méthode est d'exploiter le principe de l'apprentissage multi-tâches. Cela consiste à entraîner un modèle avec plusieurs tâches en même temps afin que les compétences obtenues avec chaque tâche participent à améliorer les performances globales du modèle. Cette méthode est ainsi particulièrement adaptée pour la détection de relations implicites associée à la prédiction de marqueurs discursifs. Liu & Strube (2023) vont utiliser cette méthode en intégrant aussi le principe de *Schedule Sampling* entre la détection de marqueurs discursifs et celle de relations implicites. D'une part, un modèle BERT est entraîné pour la détection de marqueurs discursifs par masquage du marqueur. Pour la tâche de détection de relations implicites, une paire d'arguments et un marqueur discursif sont donnés en entrée au modèle et le *Schedule Sampling* va consister à fournir de plus en plus de marqueurs discursifs prédits au fur et à mesure de l'entraînement, au lieu des marqueurs discursifs de référence. Ainsi, le modèle de prédiction de marqueurs discursifs va enrichir les informations du modèle de relation implicite et l'alternance entre les données prédites et les données de référence réduit les erreurs possibles avec les marqueurs discursifs obtenus automatiquement. La meilleure exactitude obtenue est de 76,23 (F-mesure : 71,15) pour la classification des quatre relations de premier niveau du PDTB 3.0 et de 65,51 (F-mesure : 54,92) pour la classification de 14 relations de deuxième niveau du PDTB 3.0.

En outre, une autre manière d'améliorer la détection de relations implicites est d'apprendre au modèle la correspondance entre un marqueur et une relation du discours. Ainsi, en effectuant une tâche de prédiction de marqueurs discursifs, les résultats obtenus vont être exploités par le modèle pour inférer la relation attendue. Wu *et al.* (2023) vont en particulier utiliser le principe de distillation de connaissance en implémentant un modèle enseignant-élève. Le modèle enseignant génère des probabilités de prédiction pour chaque marqueur discursif. De plus, il effectue ses prédictions en ayant l'information de la relation du discours associée au marqueur. Le modèle élève utilise en parallèle les connaissances du modèle enseignant en récupérant les probabilités des marqueurs discursifs et effectue la prédiction d'un marqueur. La relation du discours associée à ce marqueur est le résultat de la tâche de détection de relations implicites. Cette méthode permet ainsi d'améliorer progressivement le modèle grâce à un transfert de connaissances et le modèle obtient une exactitude de 78,56 (F-mesure : 75,52) pour la classification des relations de premier niveau de PDTB 3.0 et une exactitude de 67,84 (F-mesure : 52,16) pour les relations de deuxième niveau de PDTB 3.0.

4.3 L'identification de marqueurs pour l'évaluation de systèmes de TAL

La tâche d'identification de marqueurs discursifs peut servir à évaluer certaines tâches de TAL. La prédiction de marqueurs discursifs peut ainsi être utilisée pour estimer les performances d'un analyseur discursif. À ce titre, Scholman *et al.* (2021) ont évalué les performances de quatre analyseurs discursifs dans l'identification de marqueurs discursifs. En particulier, ils ont souhaité vérifier si ces analyseurs maintenaient leurs performances sur des données de différentes disciplines. À l'exception de l'analyseur discursif Discopy de Knaebel & Stede (2020) dont l'architecture se base sur les réseaux de neurones, les analyseurs discursifs de Lin *et al.* (2014), Nie *et al.* (2019) et Lan *et al.* (2017) s'appuient principalement sur des caractéristiques syntaxiques. L'évaluation de leur performance s'est faite avec une tâche de détection de marqueurs discursifs avec les données du PDTB, du TED-MDB qui représente de l'oral préparé, DiscoSpice (Rehbein *et al.*, 2016) qui est de l'oral spontané et des données médicales (Prasad *et al.*, 2011). Les résultats ont alors montré que Discopy, l'analyseur discursif qui s'appuie sur des plongements contextualisés, a globalement de meilleurs résultats sur les différents corpus. En outre, Bourgonje & Lin (2024) ont présenté une procédure d'analyse discursive pour le chinois, l'italien, le portugais, le turc et le thaï et l'ont évaluée avec une tâche de détection de marqueurs discursifs. Cette procédure s'appuie sur la traduction en anglais des textes d'une des langues mentionnées, l'annotation de la traduction anglaise par un analyseur adapté pour l'anglais et la projection de ces annotations sur le texte initial. La traduction et l'alignement sont ainsi utilisés pour annoter les langues peu dotées en données discursives. En comparaison à un outil tel que DisCut de Metheniti *et al.* (2023) qui est un analyseur discursif initialement multilingue et qui a une architecture Transformer, la pipeline de Bourgonje & Lin (2024) a des performances moindres dans toutes les langues, ce qui semble indiquer que les outils de traduction et d'alignement peuvent être améliorés. De plus, on pourrait aussi penser que les spécificités linguistiques des marqueurs discursifs changent plus ou moins d'une langue à une autre et la traduction ne rendrait pas compte assez justement de ces subtilités.

En outre, pour les différentes tâches en rapport avec les marqueurs discursifs telles que la prédiction ou l'implémentation d'analyseurs discursifs, l'architecture Transformer est souvent employée pour ses performances satisfaisantes dans un grand nombre de tâches de traitement automatique des langues. Cette architecture probabiliste représente pertinemment les particularités de la langue mais ce type de modèle est une boîte noire, ce qui complexifie l'interprétation de son fonctionnement. Ainsi, différentes tentatives sont effectuées pour tenter de comprendre comment les aspects linguistiques du discours sont représentés dans ces types de modèles. Dans ce sens, Pandia *et al.* (2021) étudient le comportement des modèles basés sur *Transformer* par le biais des marqueurs discursifs. En particulier, ils étudient leurs compétences en pragmatique à travers des paires de phrases où le modèle est censé prédire le marqueur discursif adéquat entre les phrases. Par exemple, entre des marqueurs discursifs concessifs et causaux, il est attendu du modèle de choisir un marqueur discursif en lien avec la relation. Un autre test consiste à capturer la capacité du modèle à identifier les usages temporels du marqueur *and*. Ces différents tests psycholinguistiques qui mettent en avant les marqueurs discursifs ont ainsi permis de montrer que les modèles BERT, RoBERTa et AlBERTa ont des difficultés à capturer les subtilités pragmatiques et le concept de temporalité. Par ailleurs, Cong *et al.* (2023) analysent la représentation des marqueurs discursifs *however* et *even so* au sein des modèles GPT-2 Base, DistilGPT-2 et GPT-Neo. Un ensemble de quatre types de phrases cohérentes et incohérentes comportant ces marqueurs discursifs est donné en entrée à ces modèles et le score de surprise obtenu pour le marqueur est analysé. Le score de surprise est le logarithme négatif de la probabilité d'un token dans le modèle et plus ce score est élevé, plus on considère que ce token est improbable et peut être considéré comme incohérent pour le modèle. Dans ce cas de figure, l'identification des

marqueurs discursifs prend la forme d'une probabilité attribuée par un modèle auto-régressif. Par ailleurs, ces différents tests ont montré que les larges modèles de langue ont un fonctionnement qui se rapproche des humains pour le marqueur discursif *even so* et il semblerait aussi que les particularités de ce marqueur discursif concessif sont représentées par le modèle.

5 L'utilisation de la sémantique des marqueurs du discours

Au-delà de l'utilisation des marqueurs discursifs pour l'amélioration de la détection de relations implicites, les marqueurs peuvent aussi être employés pour venir en support à d'autres tâches de TAL. En particulier, le sens des marqueurs discursifs peut être utilisé de manière subsidiaire.

Ainsi, Braud & Denis (2016) ont implémenté un modèle de plongements de mots anglais basé sur une approche distributionnelle, pour une tâche de détection de relations implicites, et ce modèle construit une représentation des mots par rapport aux marqueurs discursifs. En effet, ils considèrent que les mots proches d'un marqueur discursif ont une similarité rhétorique. Des méthodes statistiques, le TF-IDF, le PPMI-IDF et la fréquence, sont employées pour représenter la relation d'un mot aux 100 marqueurs discursifs du PDTB. Le corpus Bllip [5] est ensuite utilisé pour constituer un vocabulaire exhaustif de plongements lexicaux, utilisé pour représenter les données utilisées pour la tâche de détection de relations implicites. Ainsi, cette modélisation des mots permet de mettre en avant leurs propriétés discursives en représentant la similarité des mots à partir des marqueurs discursifs.

Ein-Dor *et al.* (2022) effectuent quant à eux une tâche de classification de sentiments par affinage de BERT. Afin d'enrichir les données d'entraînement, ils sélectionnent 11 marqueurs du discours qu'ils associent à une polarité positive, négative ou neutre. Une extraction de phrases commençant par un des marqueurs (suivi d'une virgule) est effectuée dans un grand corpus de presse : ces phrases sont alors étiquetées avec le sentiment associé au marqueur. Cet ensemble de phrases (sans le marqueur) obtenu est considéré comme un jeu de données faiblement étiqueté dans la mesure où les informations obtenues tendent à être bruitées car elles sont constituées sans supervision. Néanmoins, cette méthode permet d'obtenir des données annotées de manière automatique et simple. Le sens des marqueurs est ainsi exploité pour construire un jeu de données pour une tâche de classification de sentiments. Par ailleurs, ces données ont été utilisées pour l'affinage de BERT et BERT-tiny et les résultats ont indiqué que l'ajout de ce jeu de données améliore les performances des modèles de petite taille.

Enfin, Lei & Huang (2024) proposent une tâche de construction d'arbres logiques pour la classification de faux raisonnements par les modèles de langue. Ils utilisent alors des marqueurs et des relations du discours du PDTB pour former les arbres. En effet, à partir de marqueurs discursifs exprimant un sens logique, 10 relations du discours sont sélectionnées du PDTB et elles constituent des noeuds non terminaux tandis que les arguments de relation constituent les noeuds terminaux. Une requête, indiquant les différents types de raisonnement fallacieux, et un texte fallacieux sont donnés en entrée à un modèle de langue (GPT, RoBERTa, Flan, Llama) qui va construire un arbre logique et envoyer en sortie un type de raisonnement fallacieux. De même que pour Ein-Dor *et al.* (2022), le sens des marqueurs discursifs est exploité pour enrichir la structure d'un modèle. Dans le cas de Ein-Dor *et al.* (2022), il est utilisé pour obtenir des données supplémentaires pour entraîner un modèle tandis que Lei & Huang (2024) construisent un arbre logique à l'aide des informations sémantiques associées aux marqueurs discursifs.

5. https://catalog.ldc.upenn.edu/LDC2000T43

6 Conclusion

Les marqueurs discursifs font l'objet de plusieurs études en traitement automatique des langues. À travers des lexiques et des corpus annotés, des travaux ont été effectués pour recenser les marqueurs et leurs usages en langue. Cela a ainsi donné naissance à des projets d'envergure tels que DiMLex et le PDTB. Dans un premier temps, l'accent a été mis sur des langues bien dotées telles que l'anglais, l'allemand ou encore le français, puis ces ressources ont été généralisées dans des langues moins voire peu dotées. En particulier, dans les langues peu représentées en TAL, les chercheurs ont généralement recours à des corpus parallèles et des méthodes d'alignement pour constituer des ressources lexicales.

À partir des ressources qui permettent d'identifier les marqueurs discursifs et leurs usages, plusieurs recherches se sont fixées comme objectif d'améliorer l'identification des marqueurs du discours. Les modèles basés sur l'architecture *Transformer* permettent d'obtenir des plongements lexicaux contextualisés et contribuent à de meilleurs résultats dans cette tâche. En outre, la prédiction des marqueurs discursifs est souvent intégrée à des modèles de détection de relations implicites ou de plongements lexicaux afin de les améliorer. La méthode la plus communément employée est l'affinage de modèles de langue pré-entraînés en y intégrant la tâche de prédiction. Pour l'évaluation d'analyseurs discursifs ou celle des compétences de modèles de langue, la tâche de détection de marqueurs discursifs est pertinente pour souligner les performances d'un outil ou son fonctionnement interne. Enfin, le sens des marqueurs discursifs est utilisé pour apporter des données supplémentaires dans des tâches telles que la classification de sentiments ou celle de raisonnement fallacieux.

Ainsi, les marqueurs discursifs, en tant que connecteurs, sont des indicateurs de relations utiles pour extraire des données ou évaluer des systèmes de traitement automatique des langues. De plus, ils peuvent être exploités pour enrichir la sémantique de plongements lexicaux. Par ailleurs, l'entraînement des modèles à la prédiction de marqueurs est une stratégie pertinente pour améliorer l'identification de relations implicites dont les applications s'étendent aussi bien aux tâches de question-réponse qu'à la génération de résumés ou encore à la traduction automatique. Actuellement, ces tâches peuvent également être exécutées par les larges modèles de langues qui parviennent à des résultats très satisfaisants pour plusieurs problèmes de TAL mais nécessitent une grande quantité de ressources matérielles. On pourrait alors se demander si la sémantique des marqueurs discursifs pourrait apporter une amélioration à des systèmes aussi performants. Une autre analyse possible serait la comparaison des larges modèles de langue à des architectures moins coûteuses affinées avec une tâche d'identification de marqueurs discursifs. Une évaluation pourrait ainsi être effectuée en termes de quantité de ressources computationnelles mobilisées par rapport aux performances obtenues sur un problème donné.

Il est également important de noter que la majorité des études sur les marqueurs discursifs en TAL met l'accent sur l'usage des marqueurs dans le contexte de corpus écrits, au détriment des corpus oraux ou de leurs transcriptions. Quelques tentatives ont été menées pour étudier les marqueurs en contexte oral, à l'image de Riccardi *et al.* (2016) qui utilisent des algorithmes d'apprentissage automatique et des attributs acoustiques pour détecter des marqueurs discursifs. Silvano *et al.* (2022) ont présenté un corpus parallèle multilingue basé sur des transcriptions de conférences TED et se sont appuyés sur des éléments de la norme ISO 24617 [6] pour annoter les marqueurs propres à l'oral. Ainsi, il pourrait être intéressant d'explorer plus en profondeur les marqueurs discursifs dans des données orales mais aussi de développer des techniques adaptées pour l'analyse de marqueurs discursifs propres à l'oral.

6. https://www.iso.org/fr/standard/74577.html

Références

BOURGONJE P., HOEK J., EVERS-VERMEUL J., REDEKER G., SANDERS T. & STEDE M. (2018). Constructing a lexicon of dutch discourse connectives. *Computational Linguistics in the Netherlands Journal*, **8**, 163–175.

BOURGONJE P. & LIN P.-J. (2024). Projecting annotations for discourse relations : Connective identification for low-resource languages. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, p. 39–49.

BRAUD C. & DENIS P. (2016). Learning connective-based word representations for implicit discourse relation identification. In J. SU, K. DUH & X. CARRERAS, Éds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 203–213, Austin, Texas : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1020.

BRAUD C., LIU Y. J., METHENITI E., MULLER P., RIVIÈRE L., RUTHERFORD A. & ZELDES A. (2023). The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In C. BRAUD, Y. J. LIU, E. METHENITI, P. MULLER, L. RIVIÈRE, A. RUTHERFORD & A. ZELDES, Éds., *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, p. 1–21, Toronto, Canada : The Association for Computational Linguistics. DOI : 10.18653/v1/2023.disrpt-1.1.

BRAUD C., ZELDES A., RIVIÈRE L., LIU Y. J., MULLER P., SILEO D. & AOYAMA T. (2024). DISRPT : A multilingual, multi-domain, cross-framework benchmark for discourse processing. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 4990–5005, Torino, Italia : ELRA and ICCL.

CARON B., COURTIN M., GERDES K. & KAHANE S. (2019). A surface-syntactic UD treebank for Naija. In M. CANDITO, K. EVANG, S. OEPEN & D. SEDDAH, Éds., *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, p. 13–24, Paris, France : Association for Computational Linguistics. DOI : 10.18653/v1/W19-7803.

CHAPADOS MUERMANS T. & KOSSEIM L. (2022). A bert-based approach for multilingual discourse connective detection. In *International Conference on Applications of Natural Language to Information Systems*, p. 449–460 : Springer.

CHIARCOS C. (2022). Inducing discourse marker inventories from lexical knowledge graphs. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2401–2412, Marseille, France : European Language Resources Association.

CHIARCOS C. & IONOV M. (2021). Linking discourse marker inventories. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, p. 40–1 : Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

CONG Y., CHERSONI E., HSU Y.-Y. & BLACHE P. (2023). Investigating the effect of discourse connectives on transformer surprisal : language models understand connectives ; even so they are surprised. In *6th BlackboxNLP Workshop : Analyzing and Interpreting Neural Networks for NLP*.

CONNEAU A., KIELA D., SCHWENK H., BARRAULT L. & BORDES A. (2017). Supervised learning of universal sentence representations from natural language inference data. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*

Processing, p. 670–680, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1070.

COSTA N. F., CHENG Y., MUERMANS T. C., HANEL B. & KOSSEIM L. (2023). Automatic identification of chinese paired discourse connectives. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, p. 114–117 : IEEE.

DANLOS L., ANTOLINOS-BASSO D., BRAUD C. & ROZE C. (2012). Vers le FDTB : French Discourse Tree Bank. In G. ANTONIADIS, H. BLANCHON & G. SÉRASSET, Éds., *Actes de la conférence conjointe JEP-TALN-RECITAL, volume 2 : TALN*, volume 2 de *Actes de la conférence conjointe JEP-TALN-RECITAL, volume 2 : TALN*, p. 471–478, Grenoble, France : ATALA/AFCP. HAL : hal-00703407.

DANLOS L., COLINET M. & STEINLIN J. (2015). Fdtb1, première étape du projet «french discourse treebank» : repérage des connecteurs de discours en corpus. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (17).

DAS D., SCHEFFLER T., BOURGONJE P. & STEDE M. (2018). Constructing a lexicon of English discourse connectives. In K. KOMATANI, D. LITMAN, K. YU, A. PAPANGELIS, L. CAVEDON & M. NAKANO, Éds., *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, p. 360–365, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5042.

DAS D. & STEDE M. (2018). Developing the bangla rst discourse treebank. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

DAS D., STEDE M., GHOSH S. S. & CHATTERJEE L. (2020). DiMLex-Bangla : A lexicon of Bangla discourse connectives. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1097–1102, Marseille, France : European Language Resources Association.

DOSTIE G. (2004). Pragmaticalisation et marqueurs discursifs. *(No Title)*.

EIN-DOR L., SHNAYDERMAN I., SPECTOR A., DANKIN L., AHARONOV R. & SLONIM N. (2022). Fortunately, discourse markers can enhance language models for sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, p. 10608–10617.

FELTRACCO A., JEZEK E., MAGNINI B. & STEDE M. (2016). Lico : A lexicon of italian connectives. *CLiC it*, p. 141.

KISHIMOTO Y., MURAWAKI Y. & KUROHASHI S. (2020). Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1152–1158, Marseille, France : European Language Resources Association.

KNAEBEL R. & STEDE M. (2020). Contextualized embeddings for connective disambiguation in shallow discourse parsing. In C. BRAUD, C. HARDMEIER, J. J. LI, A. LOUIS & M. STRUBE, Éds., *Proceedings of the First Workshop on Computational Approaches to Discourse*, p. 65–75, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.codi-1.7.

LAN M., WANG J., WU Y., NIU Z.-Y. & WANG H. (2017). Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In M. PALMER, R.

HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1299–1308, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1134.

LASCARIDES A. & ASHER N. (2007). *Segmented Discourse Representation Theory : Dynamic Semantics With Discourse Structure*, volume 3, p. 87–124. DOI : 10.1007/978-1-4020-5958-2_5.

LEI Y. & HUANG R. (2024). Boosting logical fallacy reasoning in LLMs via logical structure tree. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 13157–13173, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.730.

LI S., WANG L., CAO Z. & LI W. (2014). Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 25–35.

LIN Z., KAN M.-Y. & NG H. T. (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In P. KOEHN & R. MIHALCEA, Éds., *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, p. 343–351, Singapore : Association for Computational Linguistics.

LIN Z., NG H. T. & KAN M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, **20**(2), 151–184.

LIU W., FAN Y. & STRUBE M. (2023). HITS at DISRPT 2023 : Discourse segmentation, connective detection, and relation classification. In C. BRAUD, Y. J. LIU, E. METHENITI, P. MULLER, L. RIVIÈRE, A. RUTHERFORD & A. ZELDES, Éds., *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, p. 43–49, Toronto, Canada : The Association for Computational Linguistics. DOI : 10.18653/v1/2023.disrpt-1.4.

LIU W. & STRUBE M. (2023). Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15696–15712, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.874.

MALMI E., PIGHIN D., KRAUSE S. & KOZHEVNIKOV M. (2018). Automatic prediction of discourse connectives. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

MANN W., MATTHIESSEN C. & THOMPSON S. (1989). Rhetorical structure theory and text analysis. *Discourse Description : Diverse Linguistic Analyses of a Fund Raising Text*, p.6̃6. DOI : 10.1075/pbns.16.04man.

MARCHAL M., SCHOLMAN M. & DEMBERG V. (2021). Semi-automatic discourse annotation in a low-resource language : Developing a connective lexicon for Nigerian Pidgin. In C. BRAUD, C. HARDMEIER, J. J. LI, A. LOUIS, M. STRUBE & A. ZELDES, Éds., *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, p. 84–94, Punta Cana, Dominican Republic and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.codi-main.8.

MARCU D. & ECHIHABI A. (2002). An unsupervised approach to recognizing discourse relations. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of*

the Association for Computational Linguistics, p. 368–375, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073145.

MENDES A., DEL RÍO GAYO I., STEDE M. & DOMBEK F. (2018). A lexicon of discourse markers for portuguese - ldm-pt. In *International Conference on Language Resources and Evaluation*.

METHENITI E., BRAUD C., MULLER P. & RIVIÈRE L. (2023). DisCut and DiscReT : MELODI at DISRPT 2023. In C. BRAUD, Y. J. LIU, E. METHENITI, P. MULLER, L. RIVIÈRE, A. RUTHERFORD & A. ZELDES, Éds., *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, p. 29–42, Toronto, Canada : The Association for Computational Linguistics. DOI : 10.18653/v1/2023.disrpt-1.3.

MILTSAKAKI E., PRASAD R., JOSHI A. & WEBBER B. (2004). The Penn Discourse Treebank. In M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA & R. SILVA, Éds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

MÍROVSKÝ J., SYNKOVÁ P., RYSOVÁ M. & POLAKOVA L. (2017). Czedlex – a lexicon of czech discourse connectives. *The Prague Bulletin of Mathematical Linguistics*, **109**. DOI : 10.1515/pralin-2017-0039.

NIE A., BENNETT E. & GOODMAN N. (2019). DisSent : Learning sentence representations from explicit discourse relations. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4497–4510, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1442.

OZA U., PRASAD R., KOLACHINA S., SHARMA D. M. & JOSHI A. (2009). The hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, p. 158–161.

PANCHENKO A., RUPPERT E., FARALLI S., PONZETTO S. P. & BIEMANN C. (2018). Building a web-scale dependency-parsed corpus from CommonCrawl. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

PANDIA L., CONG Y. & ETTINGER A. (2021). Pragmatic competence of pre-trained language models through the lens of discourse connectives. In A. BISAZZA & O. ABEND, Éds., *Proceedings of the 25th Conference on Computational Natural Language Learning*, p. 367–379, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.conll-1.29.

PASCH R., BRAUSSE U., BREINDL E. & WASSNER U. H. (2003). *Handbuch der deutschen Konnektoren*, volume 1. de Gruyter Berlin.

PRASAD R., DINESH N., LEE A., MILTSAKAKI E., ROBALDO L., JOSHI A. & WEBBER B. (2008). The Penn Discourse TreeBank 2.0. In N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS & D. TAPIAS, Éds., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA).

PRASAD R., MCROY S., FRID N., JOSHI A. & YU H. (2011). The biomedical discourse relation bank. *BMC bioinformatics*, **12**, 1–18.

PRASAD R., WEBBER B. & LEE A. (2018). Discourse annotation in the PDTB : The next generation. In H. BUNT, Éd., *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, p. 87–97, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

RAUH S., ZACZYNSKA K. & BOURGONJE P. (2023). Toward a multilingual connective database : Aligning German/French concessive connectives. In M. GEORGES, A. HERYGERS, A. FRIEDRICH & B. ROTH, Éds., *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, p. 77–84, Ingolstadt, Germany : Association for Computational Linguistics.

REHBEIN I., SCHOLMAN M. & DEMBERG V. (2016). Annotating discourse relations in spoken language : A comparison of the PDTB and CCR frameworks. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1039–1046, Portorož, Slovenia : European Language Resources Association (ELRA).

RICCARDI G., STEPANOV E. A. & CHOWDHURY S. A. (2016). Discourse connective detection in spoken conversations. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6095–6099 : IEEE.

ROZE C., DANLOS L. & MULLER P. (2012). Lexconn : a french lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (10).

RYSOVÁ M., SYNKOVÁ P., MÍROVSKÝ J., HAJIČOVÁ E., NEDOLUZHKO A., OCELÁK R., PERGLER J., POLÁKOVÁ L., SCHELLER V., ZDEŇKOVÁ J. & ZIKÁNOVÁ Š. (2016). Prague discourse treebank 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

SCHIFFRIN D. (1987). *Discourse markers*. Volume 5. Cambridge University Press.

SCHOLMAN M., DONG T., YUNG F. & DEMBERG V. (2021). Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, p. 95–106.

SILEO D., VAN DE CRUYS T., PRADEL C. & MULLER P. (2019). Mining discourse markers for unsupervised sentence representation learning. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3477–3486, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1351.

SILVANO P., DAMOVA M., OLEŠKEVIČIENĖ G. V., LIEBESKIND C., CHIARCOS C., TRAJANOV D., TRUICĂ C.-O., APOSTOL E.-S. & BACZKOWSKA A. (2022). ISO-based annotated multilingual parallel corpus for discourse markers. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2739–2749, Marseille, France : European Language Resources Association.

SPORLEDER C. & LASCARIDES A. (2008). Using automatically labelled examples to classify rhetorical relations : An assessment. *Natural Language Engineering*, **14**(3), 369–416.

STEDE M. & UMBACH C. (1998). Dimlex : A lexicon of discourse markers for text generation and understanding. In *COLING 1998 Volume 2 : The 17th International Conference on Computational Linguistics*.

WU C., SU J., CHEN Y. & SHI X. (2019). Boosting implicit discourse relation recognition with connective-based word embeddings. *Neurocomputing*, **369**, 39–49.

WU H., ZHOU H., LAN M., WU Y. & ZHANG Y. (2023). Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5908–5923.

ZELDES A., DAS D., MAZIERO E. G., ANTONIO J. & IRUSKIETA M. (2019). The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In A. ZELDES, D. DAS, E. M. GALANI, J. D. ANTONIO & M. IRUSKIETA, Éds., *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, p. 97–104, Minneapolis, MN : Association for Computational Linguistics. DOI : 10.18653/v1/W19-2713.

ZELDES A., LIU Y. J., IRUSKIETA M., MULLER P., BRAUD C. & BADENE S. (2021). The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In A. ZELDES, Y. J. LIU, M. IRUSKIETA, P. MULLER, C. BRAUD & S. BADENE, Éds., *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, p. 1–12, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.disrpt-1.1.

ZEYREK D. & BAŞIBÜYÜK K. (2019). TCL - a lexicon of Turkish discourse connectives. In N. XUE, W. CROFT, J. HAJIC, C.-R. HUANG, S. OEPEN, M. PALMER & J. PUSTEJOVKSY, Éds., *Proceedings of the First International Workshop on Designing Meaning Representations*, p. 73–81, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3308.

ZEYREK D. & KURFALI M. (2017). TDB 1.1 : Extensions on Turkish discourse bank. In N. SCHNEIDER & N. XUE, Éds., *Proceedings of the 11th Linguistic Annotation Workshop*, p. 76–81, Valencia, Spain : Association for Computational Linguistics. DOI : 10.18653/v1/W17-0809.

ZEYREK D., MENDES A. & KURFALI M. (2018). Multilingual extension of PDTB-style annotation : The case of TED multilingual discourse bank. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

ZHOU Y. & XUE N. (2015). The chinese discourse treebank : A chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, **49**, 397–431.

ZHU Y., KIROS R., ZEMEL R., SALAKHUTDINOV R., URTASUN R., TORRALBA A. & FIDLER S. (2015). Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 19–27. DOI : 10.1109/ICCV.2015.11.