Méthodologie pour le FDTB (French Discourse Tree Bank)

Laurence Danlos

Méthodologie pour le FDTB (French Discourse Tree Bank)

Dans l'idée de disposer de corpus annotés pour le français, nous avons l'objectif de développer le FDTB (French Discourse Tree Bank), un corpus annoté pour l'analyse discursive. Le FDTB s'inspire du PDTB (Penn Discourse Tree Bank, (PDTB Group, 2008)) qui ajoute une couche d'annotation discursive (manuelle) sur le PTB-v2 (Penn Tree Bank, (Marcus *et al.*, 1999)), corpus anglais tiré du *Wall Street Journal* annoté manuellement pour la morpho-syntaxe. De même, le FDTB ajoute une couche d'annotation discursive sur le FTB (French Tree Bank, (Abeillé *et al.*, 2003)), corpus français tiré du journal *Le Monde* annoté manuellement pour la morpho-syntaxe.

La méthodologie utilisée dans le PDTB consiste à annoter les arguments de certains connecteurs et la (ou les) relations de discours exprimée(s) par ces connecteurs [1]. L'objectif du PDTB n'est donc pas d'obtenir une analyse discursive complète du texte comme celle visée en RST (Mann et Thompson, 1987) ou SDRT (Asher et Lascarides, 2003) où un graphe discursif connexe couvre **tous** les segments du texte (au même titre qu'une analyse syntaxique d'une phrase couvre **tous** les mots de la phrase). A rebours, notre objectif est d'obtenir une couverture totale car seule une analyse complète d'un texte permet de rendre compte de sa cohérence et d'en extraire des informations ou de le résumer adéquatement, par exemple. Notre méthodologie est donc hybride entre celle employée dans le PDTB et celle employée dans les corpus annotés selon RST ou SDRT, voir le corpus français Annodis (Péry Woodley *et al.*, 2009).

Plus précisément, nous voulons dans une première étape annoter **tous** les connecteurs de discours (explicites ou implicites) et par là-même segmenter le texte en EDU (Elementary Discourse Unit). La segmentation en EDU est aussi la première étape qui a été effectuée pour Annodis mais sans s'appuyer sur les connecteurs de discours et sans s'appuyer sur une analyse syntaxique (manuelle ou non) du corpus. A l'inverse, nous nous appuyons sur l'analyse syntaxique en particulier pour repérer les connecteurs implicites inter-phrastiques (voir ci-dessous).

Cette première étape débouchera sur un corpus partiellement annoté qui sera librement distribué et qui peut déjà rendre des services à la commnunauté, par exemple pour repérer les différents emplois d'un connecteur donné. Elle se divise en deux sous-étapes, repérer les connecteurs explicites puis les connecteurs implicites, qui sont brièvement décrites ci-dessous.

Il restera ensuite, d'une part, à annoter les argument des connecteurs. Un argument, qui peut être discontinu, peut être un EDU ou une unité complexe composée minimalement de deux EDUS et un connecteur. D'autre part, à indiquer la (ou les) relations de discours exprimée(s) par les connecteurs, en s'appuyant sur une hiérarchie arborescente des relations de discours qui différent quelque peu de celle utilisée dans le PDTB (Danlos et Roze, 2011).

Identification des connecteurs explicites : Le lexique LEXCONN répertorie une liste aussi exhaustive que possible des connecteurs du français : il comporte plus de 300 éléments (Roze

1. Les connecteurs annotés sont les connecteurs explicites appartenant à une liste d'une centaine d'éléments, plus certains connecteurs implicites mais pas tous (PDTB Group, 2008). D'autres informations comme la source des arguments et des relations de discours sont aussi annotées, mais nous ne les décrirons pas ici.

et al., 2012). Le FDTB va annoter tous les connecteurs de LEXCONN. Le problème posé est celui de la désambiguation entre emplois discursifs versus non discursifs. En effet, un même mot ou groupe de mots peut avoir un emploi comme connecteur de discours et un emploi non discursif, comme illustré pour *à ce moment-là* en (1) : en (1a), il s'agit d'un emploi discursif (souligné) mais pas en (1b), (Roze *et al.*, 2012).

(1) a. Tu as l'air de penser qu'elle n'est pas honnête. A ce moment-là, ne lui raconte rien.
b. Il a commencé à pleuvoir. Marie est arrivée à ce moment-là.

Identification des connecteurs implicites : Les connecteurs implicites (notés Ø) interphrastiques apparaissent dans des phrases complexes comportant par exemple le signe de ponctuation " :", (2a), un gérondif, (2b), ou une relative explicative, (2c). L'analyse syntaxique permet de développer un outil de pré-annotation qui demande cependant une révision manuelle (Antolinos-Basso, 2012). L'identification d'éventuels connecteurs (explicites ou implicites) à l'intérieur d'une phrase permet de la segmenter en EDU (une phrase sans connecteur constitue un EDU).

(2) a. Luc a quitté la réunion : Ø il était fatigué.
b. Luc a fait la vaiselle Ø en chantant.
c. Luc, Ø qui était fatiqué, a quitté la réunion.

Un connecteur implicite est ajouté à l'initiale d'une phrase lorsque celle-ci ne comporte pas de connecteur explicite comme en (1b) répété en (3a). Il faut faire attention au fait qu'un connecteur ne se trouve pas forcément à l'initiale d'une phrase : il peut être au milieu du noyau verbal, (3b), ou même dans une phrase enchâssée, (3c).

(3) a. Il a commencé à pleuvoir. Ø Marie est arrivée à ce moment-là.
b. Luc est allé à Dax. Il est ensuite allé à Pau.
c. Luc est allé à Dax. Jane croit qu'ensuite il est allé à Pau.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer Academic Publishers, Dordretch.

Antolinos-Basso, D. (2012). Les connecteurs implicites dans le FDTB. Mémoire de Master, Université Paris Diderot.

Asher, N. et Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge.

Danlos, L. et Roze, C. (2011). Hiérarchie des relations de discours dans le FDTB. Rapport technique, ALPAGE, Université Paris Diderot.

Mann, W. et Thompson, S. (1987). Rhetorical structure theory. *In* Kempen, G., éditeur : *Natural Language Generation*, pages 85–95. Martinus Nijhoff Publisher, Dordrecht.

Marcus, M., Santorini, B., Marcinkiewicz, M. A. et Taylor, A. (1999). Building a treebank for english. *In Treebank-3*. Linguistic Data Consortium, Philadelphie.

PDTB Group (2008). The Penn Discourse Treebank 2.0 annotation manual. Rapport technique, Institute for Research in Cognitive Science, University of Philadelphia.

Péry Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho Dac, L.-M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez Couret, M., Vieu, L. et Widlöcher, A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. *In Proceedings of TALN 2009*, pages 190–196, Senlis, France.

Roze, C., Danlos, L. et Muller, P. (2012). LEXCONN : a French lexicon of discourse connectives. *Discours*, 10.