7 Conclusion

Dans cet article, nous avons proposé une méthode d'apprentissage pour la segmentation de textes arabes en unités de discours minimales. Cette méthode prédit également les UDM imbriqués. À notre connaissance il s'agit du premier travail qui s'adresse directement à la segmentation du discours en langue arabe. En effet, le seul travail existant tend à produire un discours Treebank arabe (Al-Saif et Markert, 2010) qui étend le discours Penn Treebank (PDTB) pour l'arabe standard moderne (MSA). Dans ce corpus, les éléments annotés sont les connecteurs de discours et leurs relations signalées et non pas la structure discursive complète du texte. Nous avons proposé une approche multi-classe d'apprentissage supervisé qui prédit les frontières des UDM et non seulement les connecteurs de discours. Notre approche utilise un lexique riche (avec 174 connecteurs) et s'appuie sur une combinaison de caractéristiques typologiques, lexicales et morphologiques. Cette approche a les avantages suivants : 1) détecter les frontières des UDM même en cas d'absence de marqueurs du discours (c'est-à-dire, dans le cas des relations implicites, ce qui représentent 15% des cas dans nos corpus). 2) La prise en compte d'UDM emboitée pendant la phase de segmentation.

La segmentation du discours est la première étape vers l'analyse du discours. Une annotation des documents TES et ATB avec des relations de discours dans le cadre de la SDRT est actuellement en cours.

Références

AFANTENOS, S. D., DENIS, P., MULLER, P. et DANLOS, L. (2010). Learning recursive segments for discourse parsing. *In Proceedings of the International Conference on Language Resources and Evaluation*, (LREC 2010), Valletta, Malta

AFANTENOS, S., ASHER, N., BENAMARA, F., BRAS, M., FABRE, C., HO-DAC, M., DRAOULEC, A. L., MULLER, P., PERY-WOODLEY, M.-P., PREVOT, L., REBEYROLLES, J., TANGUY, L., VERGEZ-COURET, M. et VIEU, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*

AL-SAIF, A. et MARKERT, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic, *In Proceedings of the International Conference on Language Resources and Evaluation,* (LREC 2010), Valletta, Malta

AL-SAIF, A. et MARKERT, K. (2011). Modelling Discourse Relations for Arabic. *The proceedings of Empirical Methods in Natural Language Processing*, (EMNLP 2011), Edinburgh.

ASHER, N. et LASCARIDES, A. 2003. Logics of Conversation. Cambridge University Press.

BELGUITH, H. L., BACCOUR, L. et MOURAD, G. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. *12th Conference on Natural Language Processing (TALN'2005)*, Dourdan.

BERGER, S., PIETRA D. et DELLA V. (1996). A maximum entropy approach to natural language processing. Computational Linguistics, 22(1), 39–71.

CARLSON, L., MARCU, D., et OKUROWSKI, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *In Jan van Kuppevelt and Ronnie Smith, editors,*

Current Directions in Discourse and Dialogue. Kluwer, Dordrecht.

DA CUNHA, I., SANJUAN, E. et TORRES M. (2010). Discourse segmentation for Spanish based on shallow parsing. In *Proc. of the 9th Mexican international conference on Advances in artificial intelligence, (MICAI 2010),* 13-23. Springer-Verlag.

DEBILI, F., ACHOUR, H. et SOUISSI, E. (2002). La langue arabe et l'ordinateur, de l'étiquetage grammatical à la voyellation automatique. Correspondances n° 71 July 2002.

FISHER, S. et ROARK, B. (2007). The utility of parse-derived features for automatic discourse segmentation. *In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, 488–495, Prague, Czech Republic.

HABASH, N. (2010). Introduction to Arabic Natural Language Processing. *Synthesis Lectures on Human Language Technologies, Graeme Hirst*, editor. Morgan & Claypool Publishers.

JIRAWAN, C., THANA, S., et ASANEE K. (2005). Element Discourse Unit Segmentation for Thai Discourse Cues and Syntactic Information. *The 9th National Computer Science and Engineering Conference*, 27-28 October.

KESKES, I., BENAMARA, F. et BELGUITH, H. L. (2012). Clause-based Discourse Segmentation of Arabic Texts, *The eighth international conference on Language Resources and Evaluation (LREC 2012),* Istanbul, 21-27 may 2012.

KHALIFA, I., FEKI, Z. et FARAWILA, A. (2011). Arabic Discourse Segmentation Based on Rhetorical Methods. International Journal of Electric and Computer Sciences IJECS-IJENS, Vol: 11(1).

LE THANH, H., ABEYSINGHE, G. et HUYCK, C. (2004). Generating discourse structures for written text. *In Proc. of the 20th International Conference on Computational Linguistics (COLING)*, pages 329–335, Geneva/Switzerland.

LEE, A., PRASAD, R., JOSHI, A., et WEBBER, B. (2008). Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank. *Proc. Constraints in Discourse III Workshop.*

LÜNGEN, H., LOBIN, H., BÄRENFÄNGER, M., HILBERT, M. et PUSKAS, C. (2006). Text parsing of a complex genre. In Bob Martens and Milena Dobreva, editors, Proc*. of the Conference on Electronic Publishing (ELPUB 2006)*, Bansko, Bulgaria.

MAAMOURI, M., BIES, A., KULICK, S. KROUMA, S., GADDECHE et ZAGHOUANI, W. (2010b). Arabic Treebank (ATB): Part 3 Version 3.2. Linguistic Data Consortium, Catalog No.: LDC2010T08.

MAAMOURI, M., GRAFF, D., BOUZIRI, B., KROUNA, S., BIES, A. et KULICK, S. (2010a). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium, Catalog No.: LDC2010L01.

MANN, W.C. et THOMPSON, S. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3). 243-281.

PRASAD, A., MILTSAKAKI, R., DINESH, E., LEE, N., JOSHI, A. et WEBBER, (2008). The Penn discourse treebank 2.0. *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

SORICUT, R. et MARCU, D. (2003). Sentence level discourse parsing using syntactic and lexical

information. In *HLT/NAACL*, Edmonton, Canada.

SPORLEDER, C. et LAPATA, M. (2005). Discourse chunking and its application to sentence compression. *In Proc. of the HLT/EMNLP Conference*, Vancouver, 257–264.

SUBBA, R. et DI EUGENIO, B. (2007). Automatic discourse segmentation using neural networks. *In Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, Trento, Italy.

SUMITA, K., ONO, K., CHINO, T., UKITA, T. et AMANO, S. (1992). A discourse structure analyzer for Japanese text. *In Proceedings of the international conference on fifth generation computer systems,* Tokyo, Japan ,1133–1140.

TOFILOSKI, M., BROOKE, J. et TABOADA, M. (2009). A syntactic and lexical-based discourse segmenter. *In Proceedings of the ACL-IJCNLP 2009 Conference*, 77–80, Suntec, Singapore.

TOUIR, A., MATHKOUR, H. et AL-SANEA, W. (2008). Semantic-Based Segmentation of Arabic Texts. Information Technology Journal. Vol: 7(7).

WOLF, F. et GIBSON, E. (2006). Coherence in Natural Language: Data Stuctures and Applications. MIT Press.