-Title: **Towards Building Recursive and Complete Discourse Structures of UD/PARSEME corpus**

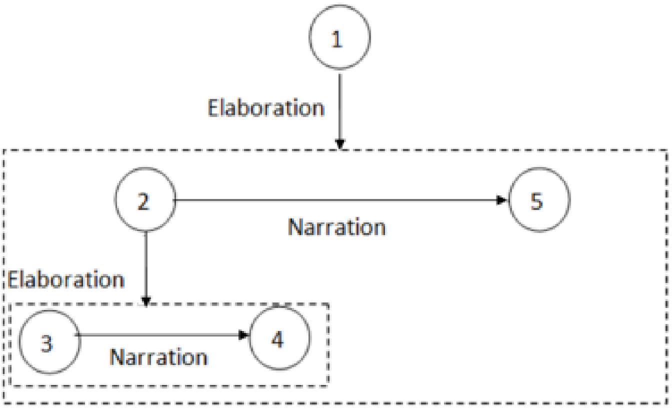
-Applicant's name and affiliation: Iskandar KESKES, Assistant professor, Faculty of Economics and Management, University of Sfax (Tunisia)

-Keywords: Arabic language, Discourse annotation, Coherent relations, Segmented Discourse Representation Theory

-Description: This project represents the initial effort towards constructing recursive and comprehensive discourse structures for the UD/PARSEME corpus. Each document will undergo annotation in accordance with the cognitive principles of the Segmented Discourse Representation Theory (SDRT) (Kamp & Reyle 1993). Within this framework, the annotation process begins by segmenting documents into elementary discourse units, which are subsequently linked by discourse relations to form complex discourse units. These complex units, in turn, may be interconnected via discourse relations to other discourse units. A document is depicted as an oriented acyclic graph that captures both explicit and implicit discourse relations, as well as complex discourse phenomena such as long-distance attachments, long-distance discourse pop-ups, and crossed dependencies. To construct such a structure, we establish a new hierarchy of relations, drawing on previous SDRT-like annotation campaigns, conducting an in-depth analysis of prior studies in rhetoric, and scrutinizing relations within the corpus. **This discourse annotation can be helpful for detecting and disambiguating multiword expressions (MWEs).**

-Example:

(1) [John had a great evening last night.]$_1$ [He had a great meal.]$_2$ [He ate salmon.]$_3$ [He devoured lots of cheese.]$_4$ [He then won a dancing competition.] $_5$



Discourse graph associated to the example.

-Explanation of how participation in the training school will be beneficial for the project:

Attending the UNIDIVE Summer School holds immense promise for the advancement of our discourse annotation project. Firstly, the specialized courses offered by renowned experts in linguistic annotation, such as Sylvain Kahane and Verginica Mititelu, directly align with our project's objectives. Kahane's course on Universal Dependencies treebank annotation and Mititelu's focus on multiword expressions offer invaluable insights and methodologies that will enrich our annotation techniques. Moreover, the opportunity to engage in hands-on activities and brainstorming sessions will foster a deeper understanding of the challenges inherent in annotating discourse structures. Collaborating with fellow trainees and instructors from diverse linguistic backgrounds will facilitate knowledge exchange and provide fresh perspectives, enhancing our project's methodology and approach.

-Open questions related to the project, which could be addressed during the brainstorming hackathon:

During the brainstorming hackathon, several key questions pertaining to our discourse annotation project could be explored. Firstly, addressing the scalability of our annotation framework to accommodate various linguistic phenomena across different languages will be crucial. This includes devising strategies to handle linguistic nuances and variations present in low-resourced languages and dialects. Additionally, discussing methods to automate certain aspects of the annotation process, particularly in identifying and disambiguating complex discourse phenomena like long-distance dependencies, will be imperative for efficiency and accuracy. Furthermore, exploring avenues to integrate machine learning and natural language processing techniques into our annotation framework could significantly enhance its scalability and adaptability. Finally, brainstorming solutions for evaluating the quality and consistency of annotated discourse structures, especially in the context of multilingual corpora, will be a pivotal aspect of ensuring the reliability and usability of our annotated data.

-Short statement of the project phase:

Our discourse annotation project is currently in the planning phase. We have laid out the foundational framework for the annotation process, including defining the scope of discourse phenomena to be annotated and establishing preliminary guidelines based on Segmented Discourse Representation Theory (SDRT) (Asher & Lascarides 2003) (Kamp & Reyle 1993). However, the actual annotation process has not yet commenced. We are in the process of finalizing the annotation protocol and assembling the necessary resources, such as linguistic expertise and annotation tools, to ensure the smooth execution of the project. Additionally, we have a previous experience on discourse relation annotation for Arabic text, where we annotated Arabic TreeBank (Keskes et al. 2014a) (Keskes et al. 2014b). This experience conduces a comprehensive review of prior studies in Arabic rhetoric and SDRT-like annotation campaigns to inform our annotation methodology. Once the planning phase is complete and all requisite preparations are in place, we will transition into the execution phase, initiating the annotation of the UD/PARSEME corpus according to our established guidelines and the SDRT framework.

A.H. Mohamed & M.R. Omer. (1999). Syntax as a Marker of Rhetorical Organization in Written Texts: Arabic and English. *International Review of Applied Linguistics in Language Teaching (IRAL)* (1999), 291-305.

Nicholas Asher & Alex Lascarides. (2003). Logics of Conversation. Cambridge University Press.

Hans Kamp & Uwe Reyle. (1993). *From Discourse to Logic.* Dordrecht.

Iskandar Keskes, Farah Benamara & Lamia Belguith Hadrich. (2014a). Learning Explicit and Implicit Arabic Discourse Relations. *Journal of King Saud University Computer and Information Sciences: Special Issue on Arabic NLP: Current State and Future Challenges. Elsevier* 26, 2 (2014), 398-416.

Iskandar Keskes, Farah Benamara & Lamia Hadrich Belguith. (2014b). Splitting Arabic Texts into Elementary Discourse Units. *ACM Trans. Asian Lang. Inf. Process.* 13, 2 (2014), 9.