Universal Discourse

1. What is the project about?

At school, we learn how words and phrases are related to each other to create the meaning of a sentence. But do sentences relate to one another to create the meaning of a whole text? Of course, we can even name some of such relations between sentences or clauses: paraphrase, explanation, comparison. We call such relations “discursive” (or simply discourse relations) because they make the whole discourse (that is, the text) coherent and comprehensible. Sometimes they are easy to identify just by looking at conjunctions or linking phrases between them but sometimes we need to think about other clues. The conjunction may be simply absent or one type of conjunction may signal various relations when used in different contexts.

For several years researchers were investigating such relations in many languages, marking them in texts. However, since there exist various ways of interpreting the linking phrases and naming the relations or roles assigned to linked phrases, the created data sets are not compatible. The lack of good data is probably the reason why large language models such as ChatGPT still have problems with identifying such relations reliably.

Our project intends to find a common denominator for the existing data sets using the model proposed by the International Organization for Standardization (ISO) to re-interpret 100,000 discourse relations and create a standard resource for further research. We called this initiative **Universal Discourse** because it seeks to bridge the fragmentation in the discourse community by combining various theories and data sets in a common one, analysing data from typologically different languages and on a large scale.

2. How will the results be achieved?

First, the common model for discourse relations will be created based on the ISO standard and using examples excerpted from existing data sets and literature. Then a selection of these data sets will be converted to the new model in a 3-step process, starting with automated conversion, through manual corrections by linguists representing languages of the data sets and finally by reconciling these interpretations by an experienced linguist.

After producing the dataset, it will be used to create discourse parsers – programs for automatic detection of discourse relations. Nowadays such programs are constructed with artificial intelligence methods – by adjusting available external language models with newly provided data.

All results of the project will be described in a book, published at the end of the 4-year project. In the meantime, partial findings will be announced to other researchers at major scientific conferences from the fields of linguistics and natural language processing.

The project will have a truly interdisciplinary nature, involving researchers with linguistic and computational backgrounds, both early-stage and experienced scholars (postdoc researchers, senior researchers, external experts and Ph.D. students) having complementary skills.

3. Which benefits will it bring to the scientific community and society?

The new, unified model of discourse relations will add a new dimension to the discussion on how various existing discourse representation frameworks are related. The data set will offer new analytic possibilities to linguists and will be a new source of valuable information for computer scientists developing tools and language models. The availability of discourse parsers will pave the way to understanding how parts of text interrelate, for example how argumentation is constructed and accomplished or which parts of text are most relevant for a generated summary. Such findings may lead to the development of new applications relevant to the society as a whole.