TextLink: Structuring Discourse in Multilingual Europe
COST Action Number IS1312

STSM Report

STSM Topic:	Polish Discourse Treebank annotation model
STSM Applicant:	Maciej Ogrodniczuk
STSM Type:	Regular (from Poland to United Kingdom)
STSM Reference Number:	COST-STSM-IS1312-30618 / ECOST-STSM-IS1312-180416-068869
STSM Dates:	From 18-04-2016 to 29-04-2016
Home Institution:	Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
Host Institution:	School of Informatics, The University of Edinburgh, UK

Introduction

The Short Term Scientific Mission had a form of two exchange visits to the Institute for Language, Cognition and Computation at University of Edinburgh in the period from 15 to 19 February 2016 and 25 to 29 April 2016 (2 weeks split into two stays in the School of Informatics together with an offline consultation period in between). The STSM activities were intended to help elaborate the annotation model of the planned Polish Discourse Treebank based on Penn Discourse Treebank.

STSM activities

During the first part of the STSM (week 1) I carried out a detailed analysis of the Penn Discourse Treebank annotation guidelines and prepared an initial version of the document describing list of relations to be used during annotation of Polish texts. I also installed and adjusted to Polish the PDTB annotation tool intended to be used in the manual annotation and started the process of annotation of the first set of texts from the Polish Coreference Corpus (http://zil.ipipan.waw.pl/PolishCoreferenceCorpus). This resource currently covers nominal direct coreference, is planned to be further extended with bridging relations and would allow for comparisons of different intra-sentence relations. During common sessions with Bonnie Webber we have discussed the model over annotation of a complete Polish sample translated to English and several individual cases from different files. I have also consulted Sam Gibbon, one of the PDTB annotators, and took part in the “hard cases” adjudicating session with Bonnie Webber, Sam Gibbon and (remotely) Alan Lee.

Between the visits I consulted the developed annotation model with Polish linguists and involved a PhD student in the annotation of sample data. This required completion of the first version of annotation guidelines which is now available for use.

During the second visit to Edinburgh (week 2) I have started annotating Polish translations of 10 TED Talks using the tools and guidelines developed in the course of the STSM. The files are intended to be used in an joint activity by partners from the UK, Germany, Portugal and Turkey. Annotation of 7 talks were completed before end of April and 3 more in May.

Main results

The following results of the STSM were accomplished during the two visits:

- the annotation/adjudication tool was successfully evaluated as compatible with the annotation of Polish texts
- a draft of annotation guidelines outlining the model of the planned Polish Discourse Treebank was prepared
- the list of senses was translated into Polish and validated both offline by a Polish linguist and in the course of annotation
- rules for segmentation of texts into elementary discourse units were drafted
- an initial lexicon of Polish discourse markers was created
- annotation of Polish samples was carried out by two annotators and hard cases were consulted.

Follow-up to STSM

The results of the STSM are intended to be included in the application for a separate Polish National Science Centre grant in HARMONIA funding scheme specifically designed for scientists carrying out research within the framework of international programmes or initiatives announced under bi- or multilateral cooperation (such as TextLink). According to the call schedule (https://www.ncn.gov.pl/finansowanie-nauki/konkursy/harmonogram) the submission start date is 15 June 2016, end date is 15 September 2016 and the funding decisions should be announced before 15 March 2017.

In March 2016 the results of the STSM were used in application for funds for continuation of the Polish branch of CLARIN ERIC research infrastructure filed to the Polish Ministry of Science and expected to be evaluated until end of June 2016. One of the tasks in this project is preparation of the discourse annotation layer in a subcorpus of the Polish Parliamentary Corpus (intended to cover 1500 samples à 300 words).

In July Deniz Zeyrek will hold a TED Talks adjudicating session in Lisbon together with Sam Gibbons, Yulia Grishina, Amália Mendes and me, provided that the funds are available for another a set of low-cost TextLink STSMs (in my case: 1–2 days accommodation only because of my visit to Lisbon for another project meeting on 6 July).