PDTB-style multilingual annotation of discourse (German, Russian)
STSM report

COST STSM Reference Number: COST-STSM-IS1312-34150
COST Applicant: Yulia Grishina, University of Potsdam
Host: Amália Mendes, University of Lisbon
Period: 21.06.2016 - 10.07.2016

July 28, 2016

The main goal of our collaborative project was a cross-lingual comparison of discourse relations in the European languages from different language groups: English (Samuel Gibbon), German and Russian (Yulia Grishina), Portuguese (Amália Mendes), Turkish (Deniz Zeyrek) and Polish (Maciej Ogrodniczuk). Our work proposal was to collaborate on the parallel discourse annotation for the languages in question using PDTB-3 annotation scheme and the development of an extended version of the annotation guidelines applicable to multiple languages.

To reach our goal, we used a sample from the TED Talks – a transcribed collection of publicly available spoken texts on different topics translated to multiple languages[1]. We agreed upon annotating 5 Talks[2] which amounts to 344 sentences in German and 341 sentences in Russian. We annotated both intra- (explicit only) and intersentential (explicit, implicit, EntRel, AltLex and NoRel) relations.

Our work was organised as follows: our annotation sessions every other day were followed by adjudication sessions on the next day, during which we discussed the resulting annotations. We used English annotations as a reference to be able to compare the annotations in different languages. These meetings were particularly helpful in order to align the annotations and identify difficult cases.

There are several issues that emerged out of the annotation and adjudication sessions:

1. Transcription: The spoken text genre poses difficulties for both the transcriber and the translators as the segmentation of continuous speech might be unclear and therefore interpreted differently in each language. For instance:

   (1) `(en)` Environment includes energy consumption, water availability, waste and pollution, just making efficient uses of resources.
   `(de)` "Umwelt" umfasst Energieverbrauch, Wassernutzung, Abfall und Umweltverschmutzung, kurz: die effiziente Nutzung von Ressourcen.

[1] https://wit3.fbk.eu
[2] Talks with the following IDs were annotated: 1927, 1971, 1976, 1978, 2009.

(pt) [Ambiente inclui consumo de energia, disponibilidade de água, lixo e poluição]$_{Arg1}$. [Trata de_ o uso eficaz de_ os recursos]$_{Arg2}$. – `conjunction`
(ru) Экология включает в себя потребление энергии, доступность воды, отходы и загрязнение, просто эффективное использование ресурсов.

In English, German and Russian, this example consists of only one sentence inside which no discourse relations can be annotated. However, in Portuguese, this example was translated as two sentences, so that it is possible to mark an implicit relation (conjunction) between them.

2. Interpretation of relations: (1) may result in translation divergences, as translators could also interpret relations between sentences in a different way and/or use different discourse devices to express them. This can be illustrated by the following English example and its German translation:

   (2) (en) [As]$_{conn}$ [I watched people who I knew, loved ones, recover from this devastation]$_{Arg1}$, [one thing that deeply troubled me was that many of the amputees in the country would not use their prostheses]$_{Arg2}$. – `temporal:synchronous`
   (de) [Ich beobachtete Menschen, die ich kannte, liebte, wie sie sich von dieser Verwüstung erholten]$_{Arg1}$, [aber[3]]$_{conn}$ [eine Sache quälte mich zutiefst, und zwar, dass viele der Amputierten in diesem Land nicht ihre Prothesen benutzten]$_{Arg2}$. – `concession:arg2-as-denier`

3. Additional sense relations: in our data, we frequently observed rhetorical questions used by the speakers in order to capture audience's attention that were subsequently answered by the speakers themselves. To process these cases, we introduced an additional implicit sense relation - namely, *Q/A*. Consider the following example:

   (3) (en) [Are investors, particularly institutional investors, engaged]$_{Arg1}$? Well, [some are, and a few are really at the vanguard]$_{Arg2}$. – `Q/A`
   (de) [Engagieren sich Investoren, insbesondere institutionelle Investoren]$_{Arg1}$? Nun, [einige tun es, und ein paar sind wirkliche Vorreiter]$_{Arg2}$. – `Q/A`

   It should be noted that we agreed on excluding *well* (and its German counterpart *nun*) from the argument span, since it has a pragmatic, not a semantic role.

4. Relations between non-adjacent sentences: According to PDTB guidelines, only relations between adjacent sentences should be annotated. From our experience, a number of relations were missed due to this limitation. For example:

   (4) (de) Wie nutzen Unternehmen ESG, um harte Geschäftsergebnisse zu erzielen? Ein Beispiel geht uns besonders zu Herzen. Im Jahr 2012 hat State Street 54 Anwendungsprogramme in die Cloud-Umgebung migriert und weitere 85 Programme eingestellt. <...> [Genial, oder]$_{Arg1}$? [Ein weiteres Beispiel ist Pentair]$_{Arg2}$. – `NoRel`

[3] *aber = but*

In such cases, we would consider appropriate to annotate an implicit relation between *[Ein Beispiel geht uns besonders zu Herzen]* and *[Ein weiteres Beispiel ist Pentair]*, since it is possible to insert a conjunction *and* between them. However, following PDTB guidelines, we had to annotate two adjacent sentences instead, marking them as *NoRel*.

In sum, we were able to achieve a considerably high level of alignment across languages and further develop and extend the annotation guidelines with respect to the spoken data and multiple languages. Furthermore, we collected the issues specific to the text genre and our languages in question and identified problematic cases which need to be addressed at a later time point. In particular, this STSM contributed to the following TextLink objectives:

- **Contribution to WG1 Resources**: we created a multilingual (English-German- Portuguese-Turkish-Polish-Russian) parallel PDTB-annotated corpus of 5 TED Talks;
- **Contribution to WG2 Annotation Guidelines**: we developed the extended PDTB guidelines applicable for the annotation of connectives in spoken texts in multiple languages.

Overall, this STSM contributed to the development of multilingual discourse-annotated resources and annotation guidelines, and it provided a solid basis for the future work on the annotation of discourse relations for each of the languages.

Acknowledgments

I thank Amália Mendes for being a very welcoming host, and all the participants in the annotation of the TED Talks in Lisbon - Samuel Gibbon, Deniz Zeyrek and Maciej Ogrodniczuk, - for making this fruitful collaboration possible. I would also like to thank Bonnie Weber and Manfred Stede for their remote participation in our discussions and sharing their vision on the controversial issues, and I am grateful to Alan Lee for his technical support.

Special thanks to Samuel Gibbon who collected the annotation issues and provided the English examples used in this report.