Report on the 1st WG1 Meeting

Prague, Oct 21-22, 2014

Compiled by Manfred Stede and Jiri Mirovsky

The primary goal of the meeting, as set forth in the invitation, was to form a final set of attributes for the tables we use to collect information on (i) corpora annotated with DSDs and (ii) inventories of DSDs. Day 1 of the workshop addressed (i), while day 2 dealt with (ii) plus a number of smaller discussion points, which will be summarized below.

The workshop started with an invited talk by Denise DiPersio (Linguistic Data Consortium, University of Pennsylvania).

Thereafter, the format of the meeting was roundtable discussion (no further prepared presentations).
Prior to the discussion of the corpus metadata, the meeting approved of the following formulation of the "mission" of the working group:

• Compile and enrich inventory of DSD resources
• Make this information publicly available
• Contribute to the improvement of existing resources
• Encourage the creation of new resources
• Encourage developers to make resources available
• => Develop TextLink guidelines for DSD resource creation and maintenance
• Along the way, synchronize with the other WGs

The discussion of the **metadata for corpora** lead to the decision to add the following fields to the schema that had been used for the pilot data collection:

• Mode
– Spoken
– Written
– Sign language
• Genre:
– journalistic, fiction, science, interactional
– Mixed vs homogeneous
– Such as: newspaper news, editorial, radio news, conversation, interview, scientific text, fiction, internet-based (twitter, chat, …) …
• Register (choose >= 1)
– Casual – semi-formal - formal / unspecified
– Spontaneous – semi-spont. - non-spontaneous / unspecified
• Text type – typically for written semi/formal
– Instructive, narrative, expository, descriptive, argumentative
• (for text corpora only) what aspects of document structure are preserved
• (A/V only) is the A/V there (aligned?) or „just“ the transcription? Is prosody annotated?
• (Speech only) unit of segmentation used

• Is the language translated (if so, from/to what language)?
• What tools are used for annotation, querying, browsing?
• Types of devices annotated
– Free text field, including „Intra“ versus „inter“
– explicit versus implicit
• Number of DSD instances {or relations} annotated
• Arguments of DSDs linked to them?
• Senses/Semantic labels annotated? yes/no
• Sample annotation (format: screenshot) for as many phenomena as possible, illustrating the breadth of annot.
• Pointers to related corpora
• Other annotation layers
– (in this vs other package)
– Anaphora
– Intonation/prosody
– Sentence morphosyntax, parse structures
– Information structure
– Other (semantic roles, modality, negation, speech acts, stance, rhetorical structure

It was decided that a corpus description should be split into two (or more) in accordance with the sets of annotations that are available for the (sub-)corpus.

The technical method of collecting the information (in the pilot: MS Excel) was decided to be Google Forms. Fields may contain multiple choice boxes or free text where appropriate. Jiri Mirovsky agreed to prepare the forms document in accordance with the suggested extensions/changes (see above) by the end of October. In addition, the Prague team agreed to compare our schema to metadata schemata used in the European CLARIN initiative, in order to ensure that common terminology be used where it is possible or appropriate.

The actual data collection is to take place in November 2014. The meeting decided not to use a wide-spread call for information but to include only the TextLink-related corpora, which means that all TextLink members will be asked to submit their information in accordance with the new schema. Also, the following resources can be added if their developers agree, and if the corpus turns out to confirm to our criteria (DSD-related annotation is present). The participants listed below agreed to contact the developers:
• PDTBs (English) Bonnie Webber
• ANNODIS (French) Liesbeth Degand
• Groningen: RST corpus (Dutch) Manfred Stede
• Utrecht/Nijmegen (Dutch) Liesbeth Degand
• Stuttgart: DIRNDL (German) Ekaterina Lapshinova-Koltunski
• Copenhagen Treebank (Danish, Iorn Korzen) Bonnie Webber
• Leeds: MSA (Katja Markert) Bonnie Webber
• Trento: LUNA (Italian) Lucie Polakova
• Finnish: Veronika
• Israel sign languages: Yael Maschler
Once the information is complete, it will be made available to the TextLink community via the website by Nicky Thrupp (in December, i.e. prior to the TextLink conference).

The discussion on **metadata for DSD inventories** did not lead to proposed changes to the table. It was found that several entries to the .xls table were in fact misplaced (software modules or corpora instead of DSD inventories) and should be removed.

The participants discussed issues of underlying **terminology** for some time, but this was for internal clarification only. Participants agreed not to put forward definitions and not to thoroughly address matters of harmonization among the resources, since both taks are being addressed by Working Group 2.

Furthermore, participants discussed the activity of WG1 during the upcoming TextLink conference in January 2015. It was agreed to hold a relatively short WG1-internal meeting dedicated to a final discussion of the Metadata schema and the means for making the information available to a wider audience. Also, this meeting should decide whether to compile a "guideline document" (see mission statement above), and if so, who would be in charge. Besides this WG-internal meeting, we intend to meet WG4 (tools) and discuss plans for incorporating our resource information into their technical infrastructure, and to present the two metadata schemata to the overall TextLink community during the conference.