COST IS1312:

TextLink: Structuring Discourse in Multilingual Europe

WORKING PLAN and BUDGET NOTES – Y1

TextLink Objectives

- The main objective of the Action is to coordinate the creation of a European portal of cross-linguistically available monolingual or parallel corpora that have been enriched and made interoperable and co-searchable through annotation of discourse-relational devices and the information they convey.

TextLink Objectives

- creating a portal that documents available tools and resources and allows data sharing that respects standards and recommendations developed in the European CLARIN project for Language Technology infrastructure;
- identifying commonalities and differences in the schemes for annotating DRDs in text and other records of language use;
- devising a common shared annotation scheme that can usefully and effectively capture valuable information about DRD tokens, that can evolve over time as it is applied to corpora across additional languages and genres;

TextLink Objectives

- encouraging and participating in the development of automated and semi-automated methods for identifying those elements in a language that serve as DRDs and for characterizing their semantic and pragmatic properties with respect to a shared annotation scheme. This may include words, phrases, morphemes that have undergone contextual modification (as in Turkish, Finnish or Arabic**),** syntactic structures (which require syntactic annotation), as well as null realisation;

TextLink Objectives

- encouraging and participating in the development of automated and semi-automated methods for rapidly annotating of new texts and other records of language use (e.g., transcripts of spoken language), since whenever large amounts of usable data become accessible to a field, it stimulates new tools, new discoveries and new applications;
- devising and sharing experimental methodologies for assessing the cognitive processing of DRDs both within and across languages, and for testing the cognitive validity of postulated semantic and pragmatic features used in the annotation scheme;

TextLink Objectives

- devising and applying methods for cross-linking and deriving information from annotated corpora across languages and genres, where the corpora may be parallel (i.e., cross-lingual or monolingual translations), comparable (on the same topic and within the same genre), or diverse;
- promoting awareness and use of the TextLink portal among stakeholders ;
- encouraging discussion with researchers working on topical and functional structure of texts and other records of language use, to understand their inter-relationship;

TextLink Objectives

- automatically monitoring use of the TextLink portal for resource access and use of its sharable annotation scheme, and assessing its impact in terms of new knowledge and new technology that it makes possible.

Working groups

- **WG1 – Resources**
  - in charge of assembling a list of existing corpora in the various languages, checking copyright issues, and developing a systematic description of each in the form of standardised metadata (to support interoperable search and facilitate comparisons between languages, genres, modes, …). The web portal administered by the Action will collect the information on those corpora, and assign appropriate meta-data, i.e. giving information on the language data contained in the corpus at stake.

Working groups

- **Milestones**
  - Updating the list of discourse-annotated corpora;
  - Designing a standardised metadata set;
  - Applying the metadata to each of the corpora;
  - Extending the list of corpora to additional data sets.
- **Deliverable**
  - Common metadata set
  - lexicons of DRDs that have been collected for various languages will be harmonised to make them interoperable. This will lead to the first multilingual comprehensive overview of DRD-relevant resources on a European scale

Working groups

- **WG2 – Interoperable Annotation Guidelines**
  - Based on a detailed study and comparison of both theoretical accounts and practical applications in already annotated corpora, this WG will work towards an interoperable conceptual framework for the multilingual annotation of the meanings conveyed by DRDs across European and non-European languages. This includes developing guidelines and recommendations for:
    - definitions of DRDs and criteria for identifying them in text;
    - an interoperable taxonomy of discourse relational meaning conveyed by DRDs, (partial) equivalences between these labels and those used by the different theories of discourse.

Working groups

- **Milestones**
  - recommending a minimal cross-linguistic set of DRDs to annotate, and a process to follow for identifying them in context, that takes advantage of what has been learned in previous corpus annotation;
  - encouraging discussion with researchers working on topical and functional structure of texts and other records of language use, to understand their inter-relationship;
  - identifying commonalities and differences in the schemes used for annotating DRDs;
  - devising a sharable annotation scheme, that can evolve over time as it is applied to corpora across additional languages and genres
  - devising and applying methods for cross-linking and deriving information from annotated corpora across languages and genres.
- **Deliverable**
  - Manual for Discourse Annotation (ISOcat proposal for discourse categories)

Working groups

- **WG3 – Assessment of Empirical and Cognitive Soundness**
  - This WG approaches DRDs from the perspectives of methodologically-sound cross-linguistic analysis, and of psycholinguistic experimentation. Within the framework of the Action “best practices” will be confronted, including results on inter-annotator agreement, compatibility of the taxonomies with psycholinguistic findings, and performance on automatic sense annotation.

Working groups

- **Milestones**
  - sharing experimental methodologies for assessing the cognitive processing of DRDs both within and across languages, and for testing the cognitive validity of postulated semantic and pragmatic features used in the annotation scheme;
  - measuring the interrater agreement of the different annotation schemes analysed in WG2, within and between languages;
  - performing automatic sense annotation Challenges.
- **Deliverable**
  - Experimental design and methods will constitute specific panels of the Training Schools.
  - Administration of open Challenges within one of the established Challenge frameworks such as those run annually by ConNL and SIGSem, where all interested parties can submit their solutions to specific DRD-related tasks, which can then be systematically compared and evaluated.
  - Raising additional funding to submit the proposed taxonomies to empirical testing.

Working groups

- **WG4 – Tools**
  - The first task of this WG is to oversee the construction and administration of the central web portal of the Action, which will provide pointers to all the resources collected by WG1, and make the findings of WG2 and WG4 readily available. The WG will make recommendations for multi-layer annotation tools that support effective annotation of DRDs and relations across languages, tools for automating parts of the process to enable more efficient use of human annotators, and tools that can use the assigned metadata to integrate, and enable search over multiple DRD annotated corpora.

Working groups

- **Milestones**
  - creating a portal that documents available tools and resources and allows data sharing that respects standards and recommendations developed in the European CLARIN project for Language Technology infrastructure;
  - participating in the development of automated and semi-automated methods for identifying those elements in a language that serve as DRDs and for characterizing their semantic and pragmatic properties with respect to a shared annotation scheme (input to WG 2);
  - participating in the development of automated and semi-automated methods for rapidly annotating new discourse corpora (input to WG2).
- **Deliverable**
  - Results will receive specific attention in the Training Schools and STSMs.

Work plan: year 1

- What did we promise?
  - Milestone: Kick-off meeting in combination with opening conference, meetings of WG1 and WG2
- What do we propose?
  - **Meetings**
    - **WG1** (sept-nov)
    - Updating the list of discourse-annotated corpora;
    - Designing a standardised metadata set;
    - Applying the metadata to each of the corpora;
      - **WG2** (nov-jan) + steering committee meeting
    - identifying commonalities and differences in the schemes used for annotating DRDs;
    - devising a sharable annotation scheme, that can evolve over time as it is applied to corpora across additional languages and genres
      - **Action kick off conference** + MC meeting (april – may)
  - **Short term scientific missions**