European Cooperation
in the field of Scientific
and Technical Research
- COST -

Brussels, 22 November 2013

COST 073/13

MEMORANDUM OF UNDERSTANDING

Subject : Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action IS1312: Structuring Discourse in Multilingual Europe (TextLink)

Delegations will find attached the Memorandum of Understanding for COST Action IS1312 as approved by the COST Committee of Senior Officials (CSO) at its 188th meeting on 14 November 2013.

MEMORANDUM OF UNDERSTANDING
For the implementation of a European Concerted Research Action designated as

COST Action IS1312
STRUCTURING DISCOURSE IN MULTILINGUAL EUROPE
(TEXTLINK)

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1. The Action will be carried out in accordance with the provisions of document COST 4114/13 “COST Action Management” and document COST 4112/13 “Rules for Participation in and Implementation of COST Activities” , or in any new document amending or replacing them, the contents of which the Parties are fully aware of.

2. 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.

3. The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 56 million in 2013 prices.

4. The Memorandum of Understanding will take effect on being accepted by at least five Parties.

5. The Memorandum of Understanding will remain in force for a period of 4 years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of section *2. Changes to a COST Action* in the document COST 4114/13.

A. ABSTRACT AND KEYWORDS

Effective discourse in any language is characterized by clear relations between sentences and coherent structure. But languages vary in how relations and structure are signalled. While monolingual dictionaries and grammars can characterise the words and sentences of a language and bilingual dictionaries can do the same between languages, there is nothing similar for discourse. For discourse, however, discourse-annotated corpora are becoming available in individual languages. The TextLink Action will facilitate European multilingualism by coordinating the nationally funded activities aimed at (1) identifying and creating a portal into such resources within Europe – including annotation tools, search tools, and discourse-annotated corpora; (2) delineating the dimensions and properties of discourse annotation across corpora; (3) organising these properties into a sharable taxonomy; (4) encouraging the use of this taxonomy in subsequent discourse annotation and in cross-lingual search and studies of devices that relate and structure discourse; and (5) promoting use of the portal, its resources and sharable taxonomy. With partners from across Europe, TextLink will unify numerous but scattered linguistic resources on discourse structure. With its resources searchable by form and/or meaning and a source of valuable correspondences, TextLink will enhance the experience and performance of human translators, lexicographers, language technology and language learners alike.

Keywords: multilingualism, discourse relational devices, portal development, resources for corpus annotation and search, contrastive linguistics

B. BACKGROUND

B.1 General background

When people communicate, whether in spoken or in written language, they produce more than just sequences of simple isolated sentences. To achieve coherence, they use expressions that indicate discourse relations between the sentences. The term ‘discourse relational devices’ (DRDs) is used here to refer to these expressions, available in all the world’s languages, including—but not restricted to—discourse markers, such as *because, but, however, I mean* or *well*, that help a speaker structure and organise discourse.

Both language learning and psycholinguistic studies confirm that mastering the expression of discourse relations is particularly challenging since they vary so much between languages. Even between two related languages, it is uncommon for DRDs to express exactly the same sense or to be used in exactly the same way. Their use is intuitive for native speakers but difficult for non-native ones, and their documentation is uneven. Existing grammars and dictionaries—even recent *wiktionaries* —fail to adequately capture their meaning, usage and translation.

What is becoming available however, are corpora annotated for discourse relational devices and structure in individual languages. The TextLink Action will facilitate European multilingualism (and technological support for discourse-level multilingualism) by coordinating the creation of a portal into annotated discourse resources across Europe, delineating the dimensions and properties of their annotation, organizing these properties into a sharable taxonomy, assisting in the development of tools that can use the taxonomy for annotation and cross-lingual search, and promoting use of the portal by stakeholders. In doing so, TextLink will enhance the experience and performance of human translators, lexicographers, language technology developers and language learners alike.

COST offers an optimal framework for networking and capacity-building in an Action in which Europe can and must take the lead because of its multilingual nature and research potential. The TextLink Action will unify numerous but scattered linguistic resources on discourse structure. It will cover a wide range of European and non-European languages, starting with Arabic, Catalan, Chinese, Czech, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Lingalá, Norwegian, Russian, Spanish, Swedish and Turkish, and include data of different genres and registers. The Action needs multiple languages for the contrastive work, belonging to different language families and from different geographical areas. It will be open for extension and adaption to still others. Bringing together teams with different theoretical and applied backgrounds, the TextLink Action will encourage the creation of an innovative unified framework for the study of DRDs in language use thus fostering mutually-enabling research. This will benefit senior and especially junior researchers, who will find comparable documentation for DRDs and recommendations for an interoperable annotation scheme applicable to their own projects and corpora.

B.2 Current state of knowledge

Discourse relational devices

DRDs have been the focus of linguistic research at least since Ducrot's (1980) ground-breaking work. In spite of their universal character and the crucial role they fulfil in the communicative function of language, their semantic and pragmatic nature is difficult to grasp, as may be witnessed by the hundreds of academic studies on DRDs published in the last twenty years on many different languages (for an overview, see Fischer 2006 ed.). Notwithstanding the impressive efforts made so far, scholars are far from reaching an agreement on such basic questions as the name of the class, the types of DRDs, their functions and their meanings (Fraser 1999, Loureda & Acin 2010). In addition, while many DRDs are described for some languages, in others their description is still in progress. Without shared frameworks for describing DRDs, results have been difficult to compare. At present, there are few bilingual descriptions of sets of DRDs, let alone multilingual descriptions. Nevertheless, a number of interesting initiatives have been undertaken recently, including the development of (e)dictionaries and lexical databases of DRDs: DiMLex for German, the DPDE for Spanish, the LEXCONN for French (Roze, Danlos & Muller 2012, Stede 2002). But research in this field has a number of shortcomings: first, such tools have been developed only for a few languages. Second, the meaning of DRDs being intrinsically context-bound, any list of DRDs out of context will inadequately represent their possible range of meanings in real uses. Third, the purposes and methods underlying the compilation of these lists diverge according to the intended applications (from descriptive adequacy to (semi-)automatic corpus extraction and semantic knowledge acquisition, cf. Alonso et al 2002, Hutchinson 2005). Finally, coming from different fields and languages, researchers have often worked in isolation, without awareness of each other’s work.

An additional strand of DRD research is more psycho-experimental in nature. This program studies human cognition by investigating the mechanisms underlying discourse coherence (Mak & Sanders 2013), including the (first language) acquisition and (on-line) processing of DRDs and the coherence relations they signal (Cain & Nash 2011, Evers-Vermeul & Sanders 2011, Sanders & Noordman 2000).

Challenges of DRDs for language learning, translation and NLP applications

DRDs are even more important in the context of language learning and translation. Studies have shown that discourse devices such as DRDs present special difficulties in second language teaching and learning (e.g. Crewe 1990, Granger & Petch-Tyson 1996, Hancock & Sanell 2010, Siepmann 2005) and translation (Baker 1993, Halverson 2004, Mason 1998, Romero Trillo 2012). These difficulties are implicitly recognized by the Common European Framework of Reference (CEFR), which includes the study of DRDs under the textual and pragmatic competences.

Translation (both human and machine) suffers from the fact that one-to-one correspondences are not frequent because many DRDs are multifunctional (e.g. Bazzanella 1999, Degand 2004, Hansen 1998, Hummel 2012, Zufferey & Cartoni 2012). Nevertheless, parallel corpora and translations can become useful tools to both identify the meanings and uses of DRDs and determine the most suitable translations (e.g. Aijmer & Simon-Vandenbergen 2003, Bosco & Bazzanella 2005).

Understanding DRDs is essential as well for Natural Language Processing (NLP) applications such as machine translation, automatic text generation, text summarisation, information retrieval. DRDs properties can be used as indicators of coherence relations between units thus improving a text's discourse representation laying at the basis of these applications (Alonso et al. 2003, Le & Abeysinghe 2003).

(Annotated) Corpora and Corpus Search Tools

Answers to questions involving the meaning, usage, or translation of DRDs can often only be found in corpora of naturally occurring language use. At the current time, a few small and medium-size corpora have been manually annotated with DRDs in Czech, Danish, Dutch, English, French, German, and other European and non-European languages (Alsaif 2012, Alsaif & Markert 2011, Buch-Kromann & Korzen 2010, Kolachina et al 2012, Oza et al 2009, Polachova et al 2012, Prasad et al 2008, Prasad et al 2011, Stede 2004, Tonelli et al 2010, van der Vliet et al 2011, Zeyrek et al 2009, 2010, Zhou & Xue 911), and the development of similarly annotated corpora in other languages and genres is underway. While different annotation schemes have been used (for reasons specific to the particular languages involved and/or to their developers' theoretical commitments), efforts are underway to facilitate inter-operability between schemata. This effort forms part of a more general activity of the International Organisation for Standardisation (ISO) that aims at improving the interoperability of language resources through the establishment of widely agreed basic annotation and representation concepts. The specific effort related to DRDs is known as the ISO Semantic Annotation Framework, Part 8: Semantic Relations in Discourse (Bunt, Prasad & Joshi 2012). This manual annotation is being used for developing automatic annotation methods, that can then be manually corrected (Alsaif 2102, Elwell & Baldridge 2008, Ghosh 2012, Jinova et al 2012, Lin 2012, Pitler & Nenkova 2009, Ramesh et al. 2012, Wellner 2009).

With respect to searching linguistic corpora, most tools are corpus-specific. These include the WebCorp search engine; and for syntactic tree-banks, Tgrep and TGrep2 for use with the Brown, Penn TreeBank, Switchboard, Negra, Chinese Treebank corpora; and *Corpus Queries* for flexible tree search of the British segment of the International Corpus of English. For searching parallel corpora, there are also multilingual concordance tools and interfaces (Tiedeman 2012). With respect to DRDs, the problem with all these search engines is that they are sentence-based, and can provide useful information only when a DRD relates parts of a single sentence, though the annotation tool GLOZZ for the French annotation project ANNODIS (Afantenos et al. 2012) is able to produce discourse level information or large patterns for DRDs, as does the PML-TQ tool (in the Prague Discourse Treebank). These tools need to be generalised to handle discourse level patterns in other frameworks. The TextLink network has the necessary expertise to perform this extension.

B.3 Reasons for the Action

The reasons for launching this Action concern both societal and scientific progress, and economic potential. Research on DRDs has been increasing in the last decades, but the lack of agreement on terminology, concepts and taxonomies makes it difficult to provide resources or applications that support key-activities in a multilingual EU such as (machine) translation, second language learning, and natural language technology. Networking and cooperation are crucial to compile an inventory of discourse annotated corpora and to standardise concepts and annotation procedures. The Action's coordination activities will result in a synergic integration of expertise and extant corpora, annotation schemes and contributions coming from different theoretical, disciplinary and language perspectives. Consensus is needed in order to enable already annotated corpora to "talk" to one another and to encourage standardisation in future DRD (annotation) work to the benefit of NLP applications. Scientific progress will be important because the Action will provide crucial input and tools for future research, both on the individual languages and cross-linguistically, and it will enhance the empirical assessment — both computational and psycholinguistic — of theoretically motivated hypotheses.

B.4 Complementarity with other research programmes

This Action is related to the European Network of e-Lexicography (ENeL; ongoing COST Action IS1305) in its endeavour to encourage standardisation of linguistic resources (albeit without a focus on grammatical words such as DRDs), and it is complementary to the aims of the ISO, more in particular, the efforts of the Technical Committee for "Language Resources" to improve the interoperability of language resources. The portal that will be created as a result of this COST Action's coordination activities will meet the standards developed in CLARIN and DARIAH. It is expected that the TextLink network will turn into a Research Project within the Europe 2020 Initiative: from coordinating existing research to producing together Research, Development and Innovation.

C. OBJECTIVES AND BENEFITS

C.1 Aim

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.

C.2 Objectives

The Action’s main focus is to coordinate the creation of the TextLink portal of inter-operable and co-searchable corpora of DRDs in all the languages of Europe and beyond, which should enable a language researcher, second-language teacher or translator to easily search for and find (i) tokens of a given DRD in a given language in order to retrieve all occurrences in context, with information on the discourse relation(s) each token realises and the genre, register and modality of the text in which it occurs; or (ii) tokens of a given DRD in a given language in order to retrieve its counterparts in other languages; or (iii) tokens of DRDs expressing a given discourse relation in one or more language corpora available through the TextLink portal. It should also allow developers of language technologies to easily find and take advantage of large discourse-annotated corpora.
TextLink Action will work towards its main aim through the coordination of achieving the following 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;

- 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;

- 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;

- 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 (Section C.5);

- encouraging discussion with researchers working on topical and functional structure of texts and other records of language use, to understand their inter-relationship;

- 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.

C.3 How networking within the Action will yield the objectives?

Achieving the aims of the TextLink Action requires engaging experts and Early Stage Researchers (ESRs) from contrastive linguistics, translation studies, foreign language pedagogy, psycholinguistics, computational linguistics and human language technology, who focus on discourse in numerous different languages. The range and complementarity of these fields and languages will ensure that the resulting resources meet the needs of its intended users. Achievement of the Action objectives is made possible through collaboration and networking of key European nationally funded research teams that have acquired highly relevant expertise in the linguistic description of DRDs, the development of discourse annotated corpora and tree-banks, and the study of psycholinguistic and/or computational processing of discourse. The group of researchers involved in the Action will make use of joint networking and communication activities foreseen by COST. Discussion of the core research issues will take place at the Plenary Action Conferences and yearly workshops. Four Working Groups will be dedicated to specific thematic issues (see Section D.2) with information exchange in thematic seminars and meetings. They will bring together researchers working on the same topic but with different theoretical and methodological backgrounds. Training Schools and Short Term Scientific Missions (STSM) will be organised to train and update ESRs and other Action members on the state of the art with respect to the topics treated mainly in WGs 2-4 and to introduce them to new findings and methods. The Final Action Conference will present the outcomes of the Action and devise plans for the future. In parallel, participants will undertake wide-scale and targeted communication activities to ensure involvement of external actors in the use of the TextLink portal.

C.4 Potential impact of the Action

TextLink Action will benefit stakeholders (professional translators, language teachers, language researchers, lexicographers and language technology developers) by coordinating and supporting the development of more discourse-annotated corpora across languages and genres; by facilitating the use of these corpora for cross-linguistic search and analysis; by enabling these corpora to be used as training tools for classifying and extracting discourse relations and recognizing their common patterns, which can then be used for developments in language technology; by aiming for deeper understanding of human cognitive processing of DRDs; and by enabling professional translators and second-language teachers and learners to identify precise naturally-occurring examples of DRDs that they want to be able to use correctly and effectively through its coordination and networking activities. The latter will assist in the design of language-learning curricula (including the CEFR for Languages), thereby enhancing the ability of linguistics research to make a unique and significant impact on multilingual communication in the European context.
In doing all this, the TextLink Action network will also make a significant difference in terms of capacity building, by directing researchers’ collaborative effort towards a shared goal. A considerable number of languages will be included from the start and attention will be given to ensuring extensibility to additional languages in the future.

C.5 Target groups/end users

Stakeholders of the TextLink Action include professional translators, language teachers, language researchers, lexicographers and language technology developers. While all but the professional translators contributed to preparing the Action proposal, academics who train professional translators were so involved.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

The Action will network experts from contrastive linguistics, translation studies, foreign language pedagogy, psycholinguistics, computational linguistics and human language technology, in order to assemble a unified resource covering most European languages, and open to non-European languages. This framework offers a common repository of information about and access to richly-annotated corpora, affording search for cross-linguistic correspondences across contexts. The most important research tasks to be coordinated are in the following areas:

Area 1: Establishing an inventory of existing discourse-annotated corpora, and those under construction, that will feed the TextLink Portal. This will require an expressive meta-data schema to use in characterizing each corpus, facilitating its inter-operability in cross-lingual (cross-corpus) search;

Area 2: Devising a conceptual framework and annotation scheme that enables as much as possible to be captured of the meanings conveyed by DRDs across European languages, in a way that supports interoperability of existing corpora and efficient annotation of corpora developed in the future. Additionally, information needs to be shared on the identification of DRDs in context, since languages diverge in the linguistic expressions that may function as DRDs. Collaboration between areas 2 and 4 will be encouraged to favour the development of semi-automatic DRDs identification and annotation tools;

Area 3: Assessing the empirical soundness and cognitive validity of the models of annotation proposed. Collaboration between areas 2 and 3 will be required in that preference will be given to annotation schemes that are cognitively plausible (as from discourse processing experiments) and/or are empirically endorsed by sound interrater agreement scores;

Area 4: Finding adapting and/or developing tools for multi-layer corpus annotation, for automating parts of the annotation process, and for searching across multi-level monolingual and parallel corpora annotated with DRDs; establishing the TextLink portal.

D.2 Scientific work plan methods and means

The bulk of the nationally funded research coordinated by this Action will be performed through four Working Groups (see below), which can work in parallel, with interconnections between the different research efforts. In addition to the Plenary Action Conferences where progress in the different Working Groups will be monitored, a number of joint Working Group meetings are planned when research is interdependent. As main vehicles of delivering the scientific work plan, the four Working Groups will be responsible for the coordination of the four research areas covering the different aspects of the multilingual discourse annotated corpus portal.

WG1 - Resources

This WG is 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. The most innovative aspect of the research performed in this WG lies in the design of the standardized meta-data for the description of the corpora. Existing corpora have been collected and developed independently from one another. It follows that the descriptive metadata diverge, hindering inter-corpus comparison. Sharing a minimal set of metadescriptive features (language, modality, genre, register, number of words, …) will help overcome this difficulty.
Main milestones of the Resources WG are:

- 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.

Results of this WG effort will be collected (and updated) in Deliverable 1. In addition, 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 (Deliverable 2). The Deliverables will be published on the Action’s website and will be publicly accessible.

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.

The main innovation is a multilingual perspective taken on the three issues. All existing approaches stem from studies on a single language. The key idea of the coordination activities of this Action, on the other hand, is to firmly establish a multilingual perspective from the beginning, separating the guidelines and recommendations into language-specific, language-family-specific, and language-neutral ones. As an example of a language specific guideline, in languages like Arabic and Turkish, it is sensible to routinely regard nominalisations as discourse units; while in others it is not. On the other hand, it seems to be a universal principle to take main clauses as discourse units in all languages.

Main milestones to make progress in this area are:

- 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.

The output of WG2 will result in Deliverable 3, in the form of a Manual for Discourse Annotation, which will be submitted to ISO, as a specific ISOcat proposal for discourse categories. Training

Schools and STSMs will heavily draw on the work performed by the teams involved. In addition to the WG meetings, the workshops organised by the ACL Special Interest Group of Annotation (SIGAnn) will be on the agenda for presentation of the research performed.

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. One instrument that will be implemented is the 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.

Main milestones are:

- 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.

In addition, the WG will specifically be in charge of raising additional funding to submit the proposed taxonomies to empirical testing. The instruments to be used here are thorough contrastive corpus studies on the one hand, and psycholinguistic experiments on the acquisition and the processing of DRDs across languages on the other hand. The main output format of this WG will be sharable experimental methodologies that can be carried out cross-linguistically and scientific publications. Experimental design and methods will constitute specific panels of the Training Schools.

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 (additional input to Deliverable 2), 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 (Deliverable 4). These results will receive specific attention in the Training Schools and STSMs. Given the variation of annotation across existing corpora, key to effective search is the provision of systematic linkages between them, enabling research with a truly multilingual perspective. In order to provide a technically viable and long-term solution, integration into technology platforms provided by the CLARIN projects will be established. Finally, and importantly, the WG will work stepwise toward the further integration of annotated resources, so that usage conditions of DRDs can be studied much more systematically and comfortably than today.

Milestones include:

- 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).

Research opportunities that will be afforded by the TextLink portal include:

- detecting ways in which language is organised at the discourse level; the discourse basis provided by the COST corpora can serve as the basis for more fine-grained analyses;
- conducting typological and cross-linguistic research into the ways in which discourse phenomena are expressed in different languages and/or different genres;
- gathering information about text quality: the features at the discourse level can be correlated to research on reading and text comprehension, to judgments about text quality, indicating various indicators of text quality;
- supporting document-level machine translation, since current systems operate in a sentence-by-sentence mode, and do not propagate information through the series of sentences that constitute texts. Such propagation can be indispensable for making

correct translation choices for words and phrases that depend on previous ones. The annotated corpora produced by the Action will enable researchers to improve their understanding of discourse entities and relations in translation and to infer new features for machine translation, this way improving overall translation quality.

To ensure the exchange of expertise between the research partners involved in a WG, emphasis will be put on short term scientific missions (STSM) for both senior and junior researchers. For the senior researchers, the STSMs are needed to make progress on the collaborative effort (implement decisions, evaluate alternatives, monitor progress, …), whereas they will serve as training support for the ESRs (data annotation, tool development, development and administration of open Challenges, experimental design, user surveys, …). The TextLink Action will constitute a unique environment for researchers, where especially ESRs will find the opportunity to collaborate and develop their expertise, thus improving the European research potential. In addition, Training Schools will give ESRs opportunities to interact with scholars within and outside of the TextLink network, both as trainers and trainees.

E. ORGANISATION

E.1 Coordination and organisation

The Action is organised in order to gain the widest possible benefits from the COST cooperation and networking opportunities.

Governance of the Action: The Action will be launched by an initial Management Committee Meeting. At this meeting the Chair and Vice-Chair of the Action will be elected along with four Working Group Leaders. This meeting will establish the specific *modus operandi* of the Action and will determine the following:

1. Policies on equality of gender representation and inclusion of early stage researchers.

2. Processes to identify participants in each of the Working Groups

3. Control mechanisms for the Action to ensure targets are met

4. Processes for determining the details of the STSMs and Training School programmes.

The Chair of the Action, the Vice-Chair, the Working Group Leaders and Co-Leaders, an STSM-Manager, a Training-School-Manager, and a Communication Manager will form a **Steering Group** for the governance of the Action. This body will meet on a virtual basis throughout the four years of the Action and physically before and after each Plenary Action Conference. The Steering Group will oversee the implementation of decisions made by the Management Committee and communication between and amongst participants in the Action. The Management Committee will meet twice a year during the Action and will be responsible for enacting the policies and processes determined at its launch meeting.

Action Coordination of Research: Three mechanisms are utilised by this Action in order to co-ordinate the research undertaken under its auspices: The Management Committee and its Chair will be in charge of the overall management. The Steering Group will give support to the Action management in order to optimise follow-up and coordination. The Working Groups will coordinate the research in the different areas of the Action.

The **Management Committee (MC)** will coordinate the Action in accordance with COST Guidelines (doc. 4159/10). In addition to normal prerogatives, it will produce a roadmap defining Steering Group role and mission.

The **Steering Group (SG)** will comprise the Management Committee Chair and Vice-Chair, together with Working Group (WG) Leaders and Co-Leaders as well as an STSM-Manager, a Training-School-Manager, and a Communication Manager. They will regularly submit activity reports to the MC. The SG's missions will include: implementing MC's policy decisions; organising and coordinating Action events (STSMs, meetings, workshops, conferences) on behalf of the MC; disseminating the Action's outcomes and managing the Website. The SG members will regularly liaise with the MC and the WGs.

- **Working Groups**: WGs will allow for planning and discussion of research. Researchers from different countries will derive practical benefits from sharing experimental results and data, such as the identification of new research topics and the launching of inter-country partnerships. WGs will provide the input for the yearly scientific meetings and STSMs.
- **Yearly scientific meetings** (Workshops and Training Schools): Work produced by WGs will be used in Workshops and will support Training Schools, passing on specialist knowledge and techniques to junior colleagues and colleagues from related disciplines. As far as possible, seminars and meetings will be held in conjunction with international conferences (in the form of panels or pre-conference workshops). International experts will be invited to participate as speakers and

reviewers, to assess the standard of the Action's work and anchor it in the scientific community. Events will be staged in different countries.

• **STSMs:** STSMs will enable researchers (especially early-stage) to undertake advanced training and exchange and strengthen partnerships. STSMs' scientific resources will be pooled during regular meetings.

• **Communication tools**: the Action's Website with the TextLink portal will serve as a central information and dissemination point during the life of the Action and beyond. The website will include a publicly available Action calendar and programme, information relating to the Action's progress and a programme for STSMs and Training Schools. It will act as an intranet to exchange information amongst COST participants and an internet to promote the Action to the wider public. It will hold project publications and findings. In the protected section, there will be a collaboration platform (mailing network, chat rooms and blogs), giving WG members an interactive space to: keep track of Action's progress; publish information, reports, minutes, guidelines, proceedings, etc. The Communication Manager will update the Website and coordinate content, with input from WGs.

The main milestones are the following:

- Year 1: Kick-off meeting in combination with opening conference, meetings of WG1 and WG2
- Year 2: Joint WG2&4 meeting, WG3 meeting, Training school conjoined with intermediate Action Conference, open TextLink Challenge, Deliverable 1 &Deliverable 2
- Year 3: Joint WG2-3-4 meeting, Training School, open TextLink Challenge, Deliverable 3
- Year 4: Joint WG3&4 meeting, Deliverable 4, Training School conjoined with Closing conference

STSMs will be carried out throughout the whole duration of the Action.

E.2 Working Groups

The bulk of the research coordinated by the Action will be conducted in four Working Groups, which will collaborate closely: WG1 - Resources; WG2 - Theory and Annotation Guidelines; WG3 - Empirical and Cognitive Soundness; WG4 - Tools. Each WG will have a Leader and a Co-Leader and will produce an annual report on its activities, which will be co-ordinated and summarised by the MC and made publicly available on the Action website. Their scientific advances will be utilised in the Training Schools and for the STSMs. The structures of the WG will be fully flexible, in order to enable other entities to join the Action. Entities need only make a request, detailing their experience, which will be approved by the SG.

E.3 Liaison and interaction with other research programmes

The Action will cooperate with CLARIN-ERIC (Common Language Resources and Technology Infrastructure - European Research Infrastructure Consortium), which is committed to establish an integrated and interoperable research infrastructure of language resources and their technologies and with DARIAH (Digital Research Infrastructure for the Arts and Humanities), which aims to enhance and support digitally enabled research across the arts and humanities. The Action will invite CLARIN and DARIAH experts to the appropriate WG meetings and will take into account the standards adopted by these two research infrastructures. Similarly, the Action will seek interaction with ISO TC 37/SC 4 (Language Resources) through participation to the Joint ISO-ACL/SIGSEM Workshops on Interoperable Semantic Annotation. The TextLink Action is also related to the COST Action ENeL (IS1305). While the two Actions share some of the objectives and means (creation of a portal, standardization of resources), their object of study diverges in that ENeL does not propose to address the multilingual challenges of linguistic categorisation and description, which is the main goal for the TextLink Action with regard to DRDs. The latter could however to a certain degree feed some of the lexicographic work. Contact between the Chairs of the two Actions is planned.

E.4 Gender balance and involvement of early-stage researchers

This COST Action will respect an appropriate gender balance in all its activities and the Management Committee will place this as a standard item on all its MC agendas. The Action will also be committed to considerably involve early-stage researchers. This item will also be placed as a standard item on all MC agendas.

The participants of this COST Action will reflect a wide European dimension, including researchers from all parts of Europe. The MC will also ensure that the Leaders and Co-Leaders of the Working Groups represent a wide range of European countries. Guidelines and policies for gender balance and for the involvement of ESRs will be put in place in the very first Management Committee meeting. These policies will be in accordance with the CSO Strategy for ESRs (COST 295/09) and will set a number of targets, such as a minimum of 20% of the annual budget for STSMs, implication of ESRs in the Training Schools both as trainer and trainee, giving priority for meeting reimbursement to ESRs, having ESRs as Leaders or Co-Leaders of WGs. Similarly, gender balance will be respected by proposing policies that guarantee that at least 40% of the Steering Group (Chair, Vice Chair, WG Leaders and Co-Leaders, STSM manager, Communication Manager, Training-School-Manager) will be of either gender; that at least 40% of STSMs will be allocated to either gender, and that gender balance is respected in Chair and Vice Chair roles. The implementation of these policies will be monitored by the Management Committee throughout this Action, and will occupy a place as a standard agenda item.

Capacity building of the ERA is an integral part of this Action. The Action will specifically address the issue of developing Early Stage Researchers through the mechanisms of Training Schools, through their involvement in STSMs and through this involvement, their contribution to Working Groups. This will enable them to not only enhance their specific research skills and profile but also enable them to develop strong networks across the ERA. The role of the STSMs and Training School Manager is fundamental to this process. It will provide mentoring for early stage researchers involved in this Action and ensure that their engagement maximises their developmental opportunities. COST funding will develop collaboration and exchange amongst PhD students and thus further build the ERA.

F. TIMETABLE

The Action will last for 4 years. Activities are regular and well distributed, in order to ensure better management of the overall Action. MC meetings will as far as possible be organised in common with SG meetings and will take place before or after Plenary Action Conferences or joint WG meetings. Whenever needed, the SG can meet virtually (video-conferencing).

Annual Monitoring Progress Reports and a Final Report will be issued, in line with the guidelines of the MC and as detailed in document COST 4159/10, thus providing appraisal of activities performed in the prior quarters.

	YEAR 1		**YEAR 2**	
	Semester 1	**Semester 2**	**Semester 1**	**Semester 2**
MC meetings	**x**	**x**	**x**	**x**
SG meetings	**x**	**x**	**x**	**x**

WG meetings	**-**	**x**	**x**	**x**
STSM		**2-4**	**3**	**3**
Training school	**-**	**-**	**x**	**-**
Open Challenge	**-**	**-**	**x**	**x**
Conferences	**x**	**-**	**x**	**-**
	YEAR 3		**YEAR 4**	
	Semester 1	**Semester 2**	**Semester 1**	**Semester 2**
MC meetings	**x**	**x**	**x**	**x**
SG meetings	**x**	**x**	**x**	**x**
WG meetings	**x**	**x**	**x**	**-**
STSM	**3**	**3**	**3**	**1-2**
Training school	**-**	**x**	**-**	**x**
Open Challenge	**x**	**x**	**-**	**-**
Conferences	**-**	**-**	**-**	**x**

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: BE, CZ, DE, DK, ES, FI, FR, HU, IT, NL, NO, SE, TR, UK. On the basis of national estimates, the economic dimension of the activities to be carried out under the Action has been estimated at 56 Million € for the total duration of the Action. This estimate is valid under the assumption that all the countries mentioned above but no other countries will participate in the Action. Any departure from this will change the total cost accordingly.

H. DISSEMINATION PLAN

H.1 Who?

The target audiences for the dissemination of the results of the Action include both academic and commercial language researchers, translators, courseware developers for second-language teaching, language technology applications developers, and scientists from neighbouring disciplines from all over Europe and worldwide. The TextLink portal of interoperable corpora that are annotated at the discourse level will be a valuable resource for any stakeholder interested in discourse structure in (multilingual) language use. The Action will be contacting European Council's policy makers with regard to the Common European Framework of Reference for Languages, but also translation schools and (commercial) translation bodies to promote and increase the benefit from the use of the TextLink portal and the multilingual DRD lexicon.

H.2 What?

a) The TextLink portal will make DRD annotated corpora of European languages available to the academic, scientific and commercial community. The format of the access will be decided upon in consultation with, among others, CLARIN EU. The results obtained in the WGs will also feed into the presentation of information in the TextLink portal.

b) The standardized discourse annotation scheme will be described in a Manual for Discourse Annotation, which will be submitted to ISO, as a specific ISOcat proposal for discourse categories.

c) The multilingual DRD Lexicon will be made available both on-line and as a print publication and made available to lexicographers (including the ENeL Action), translation schools and commercial translation bodies.

d) The public part of the Action website will provide the annual reports of the WGs, the roadmaps developed in the Action, a calendar detailing current and upcoming Action events and other documents created within or relevant to the Action.

e) The password-protected part of the Action website will be used to publish draft or working documents and, to create discussion forums for the WGs.

f) Members of the Action will submit joint papers to reviewed journals in the fields of discourse analysis, semantics and pragmatics, cognitive and functional linguistics, corpus linguistics, computational linguistics, machine translation and language technology in order to gain a broad academic audience.

g) Email invitations to the interim conference (year 2) will be sent to scientists who might be interested in participating in the Action.

h) In order to initiate further research on cross-linguistic comparison in a European context an electronic Newsletter and e-mail network will be launched. This will support the development of an interdisciplinary discourse community of linguists, corpus and computational linguists in Europe. It will be open to any interested stakeholder.

i) The Final Conference (year 4), which will follow a Training School, will be held in connection with a large conference on corpus linguistics. It will disseminate the end results of the Action across all WGs and to other stakeholders through the conference proceedings.

j) Training Schools will be held to update ESRs and other Action members on to the state of the art with respect to the topics treated mainly in WGs 2, 3 and 4 and to introduce them to new findings and methods.

k) Other potential stakeholders will be invited to participate in WG meetings.

H.3 How?

The Communication Manager will play a vital role in planning and executing the activities associated with dissemination in this Action. The MC will define the overall dissemination plan and supervise its implementation. Appointed (groups of) MC members will be in charge of organising specific dissemination activities of the Action as a whole and liaising with the relevant target groups. These activities will include:

a) The Action website: With its public and password-protected areas, it will constitute the main dissemination platform. It will be used to store all outputs and resources generated by and relevant to the COST Action.

b) The TextLink corpus portal will link the Action to potential stakeholders (professional translators, language teachers, language researchers, lexicographers and language technology developers). It will give access to monolingual or parallel corpora that have been enriched through annotation of discourse-relational devices (DRDs) and the information they convey.

c) Social Media and email will be used to distribute information, alert Action members to new content on the website and generate discussions between workshops and meetings.

d) Action participants will engage in dissemination activities within their own region and professional setting (for example, lectures, courses, workshops).