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

Brussels, 21 November 2012

IC1207

MEMORANDUM OF UNDERSTANDING

Subject :	Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action IC1207: PARSEME: PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing.

Delegations will find attached the Memorandum of Understanding for COST Action as approved by the COST Committee of Senior Officials (CSO) at its 186th meeting on 20 - 21 November 2012.

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

COST Action IC1207
PARSEME: PARSING AND MULTI-WORD EXPRESSIONS. TOWARDS LINGUISTIC PRECISION AND COMPUTATIONAL EFFICIENCY IN NATURAL LANGUAGE PROCESSING.

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 4154/11 "Rules and Procedures for Implementing COST Actions", or in any new document amending or replacing it, the contents of which the Parties are fully aware of.

2. The main objective of the Action is to increase and enhance the ICT support of the European multilingual heritage by bringing about a substantial progress in the understanding and modelling of Multi-Word Expressions (MWEs) within advanced multilingual Natural Language Processing (NLP) techniques, notably deep parsing.

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 2012 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 Chapter V of the document referred to in Point 1 above.

A. ABSTRACT AND KEYWORDS

The Action, PARSEME, aims at increasing and enhancing the support of the European multilingual heritage from Information and Communication Technologies (ICT). This general aim is addressed through improving linguistic representativeness, precision and computational efficiency of Natural Language Processing (NLP) applications. The Action focuses on the major bottleneck of these applications: Multi-Word Expressions (MWEs), i.e. sequences of words with unpredictable properties such as "to count somebody in" or "to take a haircut". A breakthrough in their modelling and processing can only result from a coordinated effort of multidisciplinary experts in different languages. COST is the most adequate framework answering this need. Fourteen European languages will be addressed from a cross-theoretical and cross-methodological perspective, necessary for coping with current fragmentation issues. Expected deliverables include enhanced language resources and tools, as well as recommendations of best practices for cutting-edge MWE-aware language models. The Action will lead to a better understanding of the nature of MWEs. It will establish a long-lasting collaboration within a multilingual network of MWE specialists. It will pave the way towards competitive next generation text processing tools which will pay greater attention to language phenomena.

Keywords: multilingualism, natural language processing, multi-word expressions, idioms, parsing

B. BACKGROUND

B.1 General background

The linguistic diversity of European countries and nations belongs to the main cultural heritage to be maintained and developed within Europe and beyond. At the same time, the competitiveness of European markets within the global economic landscape must rely on efficient information access and processing. Since information is most often available in a textual or spoken form, in particular in the fast evolving Internet, and its amount is constantly growing, support from Information and Communication Technologies (ICT) is crucial. Thus, methods for intelligent text processing have been developed for decades, resulting in an increasing number of applications such as information extraction, machine translation, question answering, automatic text summarization, sentiment and opinion mining, human-machine dialogue, etc. Such Natural Language Processing (NLP) applications face three essential challenges:

- linguistic precision of methods and results (reflecting, at least partly, the richness and creativity of human language),
- specificities of particular languages and language families,
- computational efficiency in the context of large amounts of (possibly noisy) data to be processed rapidly.

It has been shown that one of the key problems to be overcome in order to meet all of these requirements simultaneously are multi-word expressions (MWEs), i.e. sequences of words with some unpredictable properties such as "to count somebody in" or "to take a haircut". MWEs are truly a bottleneck of NLP, e.g. in machine translation, which tends to translate MWEs word by word. For instance Google wrongly translates:

- "to count Poland in" as: "compter la Pologne en" (FR), "contar con Polonia en" (ES), "contano in Polonia" (IT), "nach Polen rechnen" (DE), etc., and
- "European banks have to take a serious haircut" as: "les banques européennes ont à prendre une coupe de cheveux grave" (FR), "los bancos europeos tienen que tener un corte de pelo seria" (ES), "Banche europee devono prendere un taglio di capelli grave" (IT), "Europäische Banken haben eine ernste Haarschnitt nehmen" (DE), "evropske banke moraju da uzmu ozbiljan šišanje" (RS), etc.

These are meaningless, partly ungrammatical, literal, word by word translations. The difficulty stems from the highly heterogeneous behaviour of MWEs at the lexical, syntactic and semantic level. Since modelling language phenomena and providing efficient processing tools for their treatment prove difficult, most efforts have focused on ICT language tools dedicated to English. Taking a variety of languages into account has often been seen as an additional obstacle. This Action admits an opposite point of view. It sees Europe’s multilingualism – provided that it is considered within a coordinated framework – as the source of a better comprehension of general linguistic phenomena that are crucial to ICT multilingual language technologies. Thus, Europe’s multilingual heritage will not only be supported by the Action, but it will become an advantage over other major NLP communities, e.g. in the USA, Japan, China, etc.
In this context, some of the main challenges come from fragmentation issues. Europe, being multilingual at its heart, must make more effort to bring together NLP research across language and nation boundaries. Moreover, the multidisciplinary richness of sciences involved (linguistics, computing, statistics, psychology, etc.) demands convergence via a common, currently missing,

meeting place to discuss the handling of MWEs from various perspectives. A COST action enabling the creation of a multidisciplinary and multilingual network of experts is the most adequate framework to meet these challenges.

B.2 Current state of knowledge

NLP has made a considerable progress within the past decades. Language resources, such as annotated corpora, electronic lexicons and grammars, are being developed for an increasing number of languages. New algorithms make it possible to process very large amounts of textual data and produce pertinent results. New small and medium enterprises offer text processing technology transfer, and the end users become aware of the added value of NLP applications.
Despite these encouraging results NLP applications need further improvement. Currently most of them admit an (explicit or implicit) division of language phenomena into clear-cut levels: (i) tokens (indivisible text units, roughly words), (ii) morphology (properties of words e.g. number, gender, etc.), (iii) syntax (structural links between words, e.g. number/gender agreement), (iv) semantics (meaning of words and sentences). However, human languages frequently show a high degree of ambiguity and fuzziness with respect to this layer-oriented model. In particular, MWEs are placed on the frontier between these levels due to their idiosyncratic properties on the one hand, and their morphological, syntactic and semantic variations on the other hand. For instance, their meaning is often non-compositional as in "to take a haircut" (i.e. "to suffer a serious financial loss"), although they admit some syntactic variation similarly to many other expressions ("take/takes/have taken/has taken/took a serious/70% haircut"). Strictly layer-oriented language models fail to reflect this specificity, and thus yield erroneous text processing results (e.g. word-to-word translations of idioms).
Although the quantitative importance of MWEs is well known (they cover up to 30% of all words in human language utterances, and are much more numerous in lexicons than single words), the achievements in their formal representation and automatic processing are still largely unsatisfactory. Current research on MWEs shows that most proposals still concentrate either on creating MWE lexicons or on the automatic recognition of MWEs in text. Only few approaches address the links between MWEs and a comprehensive linguistic analysis of text. These approaches confirm that a proper treatment of MWEs increases both linguistic precision and robustness.
With respect to this state of the art, the Action will make a cutting-edge contribution by:

- **A highly contrastive methodology.** The Action will cross language boundaries by studying MWEs in different European languages. It will also compare points of view on MWEs in different linguistic and methodological frameworks.

- **Accounting for the richness of the linguistic heritage in Europe.** The Action will consider over 14 languages from all major European language families: Germanic (English, German, Norwegian, Swedish), Romance (French, Italian, Portuguese, Spanish), Slavic (Bulgarian, Czech, Polish, Serbian) and Finno-Ugric (Estonian, Hungarian).

- **Increasing the cohesion of different levels of linguistic processing.** Methods of explicit inclusion of MWEs into most levels of linguistic processing will be defined. The existing lexicons and grammars will increase their MWE coverage, and will be extended by dealing with morphology/syntax/semantics interface aspects. The strictly word-by-word processing framework will be abandoned, and the treatment of MWEs in syntactic and semantic processing will be brought into focus instead.

- **Developing methodologies for cost saving resource development.** In order to cope with the high cost of the development of language resources, the Action will put forward abstract MWE representation formalisms that can be mapped onto different linguistic frameworks.

- **Simultaneously accounting for both linguistic precision and robustness.** The Action will address both knowledge-based and data-driven approaches (cf. the following section), and make the most of their complementarity by enhancing and extending hybrid models. MWEs will play a central role in these considerations.

B.3 Reasons for the Action

Supporting Europe's linguistic heritage via linguistic precision and computational efficiency of NLP tools necessarily calls for a large panel of multilingual and interdisciplinary experts. Although many European languages are already addressed by national NLP research teams, these efforts need further reinforcement in crossing language barriers.
Moreover, the NLP community represents two major methodological trends. Symbolic (knowledge-

based) methods aim at an explicit modelling of linguistic phenomena via language resources such as lexicons and grammars. Conversely, statistical (data-driven) approaches offer mostly implicit language models automatically derived from annotated or raw text corpora. The complementarity of these two trends is visible in particular with respect to MWEs: (i) MWE-aware language resources help in high-quality corpus annotation, (ii) results of statistical text processing can be improved by taking MWE resources into account, (iii) MWE detection decreases the search space in both types of methods, etc. In this context there is a growing need of reinforcing interactions between the respective research communities.

Also, within the knowledge-based approaches different linguistic frameworks, such as Lexical Functional Grammar (LFG), Head-driven Phrase Structure Grammar (HPSG), Tree Adjoining Grammar (TAG), Combinatory Categorial Grammar (CCG), etc., are being used. They offer complementary points of view on language phenomena via fine-tuned lexicon and grammar rules. By increasing the interoperability of these precious resources, in particular with respect to MWEs, a better understanding of these phenomena will be achieved and grammar development for a variety of languages will be boosted.

Finally, the NLP community suffers from ill-balanced language representativeness. A large majority of efforts have been dedicated to English. More recently multilingualism issues are increasingly covered but many approaches tend to adapt English-oriented achievements to languages whose nature may be rather distant from English, for instance to morphologically rich languages. That jeopardises the quality of text processing, in particular in MWEs, where complex interdependencies between morphology and syntax take place. As a result, Europe’s linguistic heritage fails to be fully satisfactorily supported by ICT.

A COST Action will effectively cope with the above mentioned fragmentation issues. It will constitute a virtual added value to the nationally funded research by establishing a network of experts in the domain of MWEs. By attracting experts in both knowledge-based and data-driven methods, in all major linguistic frameworks, and in major European language families, the Action will reach optimal methodological and linguistic representativeness. This strong networking aspect will produce a leverage effect for the understanding and treatment of MWEs.

Thus, the Action will meet both economic/societal and scientific/technological needs by:

- directly accounting for the language variety in Europe,
- reinforcing the competitiveness of European markets with high-quality efficient and interoperable language processing tools,

- educating a wide public about a better use of such tools, and increasing its sensitivity to linguistic issues,
- achieving a better understanding of linguistic phenomena such as MWEs,
- transferring and perpetuating this understanding via training offered to young researchers (Master's and doctoral programmes, scientific missions),
- increasing the cohesion of the European Research Area (ERA), and its NLP community in particular; promoting a good integration of Early-Stage Researchers in the ERA.

B.4 Complementarity with other research programmes

The Action will partly build on current European projects such as FLaReNet (ECP-2007-LANG-617001), CLARIN (212230), META-NET (ICT-NoE-249119), ATLAS (CIP-ICT-PSP-250467) and CESAR (CIP-ICT-PSP-271022) dedicated to interoperability, standardisation and dissemination of linguistic resources and tools. However, the scope of these projects is significantly broadened by focusing at higher level text processing (i.e. syntactic and semantic parsing) and by addressing one of its major bottlenecks (i.e. MWEs).
Moreover, the current COST MUMIA Action (IC1002) addresses issues which partly overlap with those of the PARSEME Action as far as multilingualism and efficient information access are concerned. However, the means and scope of MUMIA are different since it primarily aims at paving the way towards new generation web search engines. Even if one of its Working Groups is dedicated to integrating and managing language resources, the main focus is on ensuring privacy and anonymity of data. This Action focuses instead on the necessity of a critical enhancement of the methodologies used to produce such resources by accounting for the nature and importance of MWEs.
Another possibly related initiative is the COST ISCH IS1006 Action dedicated to grammars of European sign languages. Since many linguistic processes are common to spoken and sign languages, an interesting insight might be brought into MWEs-related issues (e.g. compounding) from this network.

C. OBJECTIVES AND BENEFITS

C.1 Aim

The aim of the Action is increasing and enhancing the ICT support of the European multilingual heritage.

C.2 Objectives

In order to come closer to the achievement of this general aim, the central objectives of this Action are:

- to put multilingualism in focus of linguistic and technological studies,
- to establish a long-lasting collaboration of NLP experts within a cross-lingual, cross-theoretical and cross-methodological research network,
- to bridge the gap between linguistic precision and computational efficiency in NLP applications.

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

The objectives of this Action will be achieved by addressing one of the main challenges in NLP, i.e. MWEs, via a close collaboration of interdisciplinary experts. Specialists in linguistics and NLP will bring their expertise in the linguistic properties of MWEs in several European languages. A contrastive analysis of these properties will allow the identification of similarities and particularities, and thus will pave the way towards a better understanding of MWEs in general, and towards more universal methods for representing MWEs. Members of the Action will be asked to bring their previous achievements in the domain, such as language resources (lexicons, grammars and treebanks) or parsing tools. The non-exhaustive list of such resources and tools available within the Action should contain:

- large-coverage lexicons and extraction tools for compounds and named entities in Bulgarian, English, French, German, Italian, Polish, Serbian and Spanish,
- syntactic lexicons of verbs and valence dictionaries in Bulgarian, French and Polish,
- computational grammars for Bulgarian, Danish, English, French, German, Hungarian, Maltese, Norwegian, Polish, Portuguese and Spanish,
- large annotated corpora and treebanks in Bulgarian, English, French, German, Hungarian, Italian, Norwegian, Polish, Portuguese, Spanish, and Swedish,
- parallel corpora in Bulgarian/English and Italian/English,
- parsing tools for Basque, Bulgarian, Catalan, Czech, Danish, Dutch, English, French, German, modern Greek, Hungarian, Italian, Polish, Portuguese, Slovene, Spanish, Swedish, as well as Arabic, Chinese, Japanese, and Turkish,
- machine translation modules for Italian/English, Polish/English, Polish/German and Polish/Russian.

These resources and tools will be further enriched and enhanced within nationally funded projects, under the influence of new insights on MWEs stemming from the collaboration within the Action. These resources and tools represent a sufficient critical mass to enable new outcomes, such as common annotation guidelines, best practices, and new common designs of resources and tools. The role of Early-Stage Researchers will be crucial here. Their participation in Short-Term Scientific Missions will allow establishing particularly close links within groups of experts. Other standard COST Action instruments will also be used: Management Committee Meetings, Working Group Meetings, Workshops, Conferences, Training Schools, and Publications.

C.4 Potential impact of the Action

The main benefit will be to bring about a substantial progress in ICT support of the European multilingual heritage. For the scientific community the prime benefits will be to:

- increase the cohesion of the European Research Area (ERA) by creating a network of specialists dedicated to MWEs in different European languages (and thus cope with the fragmentation issues discussed in Sections B.1 and B.3),

- achieve a better understanding of the nature of MWEs in different European languages,
- achieve a breakthrough in the processing of MWEs (and thus gain advantage over the state of the art presented in Section B.2),
- increase the coverage and the robustness of existing text processing resources and tools,
- create new interoperable resources and tools to serve a large and versatile scientific community,
- set up a lasting and fruitful collaboration between major European scientific actors in the field, possibly via establishing a Special Interest Group.

For the general public the Action will pave the way for next generation text processing tools which will pay greater attention to language phenomena.

An important commercial impact is also expected. Investigating a phenomenon such as MWEs in various languages and theoretical frameworks simultaneously will bring new insights into its understanding and reinforce the convergence and interoperability of language resources including MWEs. The resulting high quality and comprehensive data will, in turn, lead to competitive NLP applications (turning Europe's multilingualism into a major advantage over other NLP communities, e.g. USA, China).

Since the impact of COST comes from concrete outcomes, clear deliverables have been foreseen:

1. Contrastive analysis of the linguistic properties of MWEs in different European languages.
2. Proposal of a common design for lexicons including both valence data and MWE data.
3. Lexical databases: possibly interoperable parsing-oriented MWE lexicons and valence dictionaries in several European languages.
4. Extensions of existing corpora and treebanks in several languages with MWE annotation levels.
5. Extensions of existing grammars for several European languages with rules dedicated to MWEs.
6. Definitions of abstract models (e.g. meta-grammars) of MWEs' properties that would: (i) capture linguistic richness of MWEs independently of particular grammatical frameworks, (ii) help reduce the cost of resource development, (iii) adapt to different languages studied.
7. Recommendations of best practices for MWE representation and treatment in parsing within different theoretical frameworks. The resulting designs should be maximally interoperable.
8. Extension of hybrid (knowledge-based and data-driven) methods for parsing MWEs.
9. Annotation guidelines for the representation of MWEs in treebanks.
10. A common publishing platform gathering initiatives in the field of MWEs and parsing.

11. Scientific publications in established conferences and journals in various domains.

C.5 Target groups/end users

The target groups of the Action are:

- Professionals in the field of NLP developing text processing tools for machine translation, information extraction, sentiment and opinion mining, computer-assisted language learning, text indexation and categorisation, web search, etc. These products will be able to gain better linguistic precision and efficiency due to the assets of the Action.

- Experienced young researchers who will gain a better integration into the European Research Area, via direct participation in the Action's Working Groups or via activities organised by the Action for a wider scientific public, e.g. workshops and Master's trainings.

- European institutions dealing with large amounts of multilingual data, who will gain more efficient and linguistically precise tools to support Europe's multilingualism.

- European citizens wishing to access and process textual information needed in everyday life via on-line tools. Their queries will be handled more easily in their mother tongues, or in other European languages, and will yield results of a better linguistic quality.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

The research tasks to be coordinated by the Action include: (i) contrastive analyses of the state of the art in different aspects of MWEs in different European languages, (ii) enrichment and enhancement of existing language resources in those languages, (iii) critical insight into methods and tools for text processing from the perspective of their MWE-awareness, (iv) development of recommendations, best practices and common designs for new – more MWE-aware and more interoperable – resources and methods.

The organisation of Working Groups is innovative in that they will cross boundaries between different languages, linguistic frameworks and methodological approaches. This principle will contribute to more universal linguistic descriptions, and will enable the inclusion of new languages, frameworks and methods during the implementation stage, if new potential partners prove interested.

D.2 Scientific work plan methods and means

Four Working Groups will share the workload.

1. **WG1** will address the **LEXICON/GRAMMAR INTERFACE**.
Lexicons are elementary language resources meant for modelling the properties of words. Increasingly many lexicons (including valence dictionaries) have been developed for contiguous MWEs such as compounds (e.g. "fine arts"), complex terms (e.g. "random access memory") and named entities (e.g. "the United Kingdom") in order to account for their idiosyncratic (i.e. unpredictable) properties. However, other types of MWEs, notably verbal ones, are hardly ever covered by these lexicons. One of the main issues here is discontinuity: verbal phrases admit insertions of external elements as in "TAKE A [serious] HAIRCUT", "COUNT [Italy, Spain, and the Wednesday game's winner] IN". Grammars offer natural means of expressing relations between distant constituents of such expressions, however MWE lexicons have frequently been constructed independently of grammar development. Thus, although these lexicons constitute very precious resources for simpler NLP tasks (such as shallow parsing, terminological extraction, etc.) they are usually hard to integrate in full-fledged text processing (including parsing).
Another problem is the high cost of MWE lexicon construction due to the quantitative importance of MWEs, which are much more numerous than single words. Some solutions to this problem come from relatively well developed techniques for automatic extraction of MWEs from corpora. However, the data sparseness (the fact that very many MWEs appear very rarely in corpora), as well as frequent MWE paraphrasing (expressing the same meaning using textually different surface forms) remain important challenges.
In this context, the objectives of WG1 will be:

- a better understanding of linguistic properties of MWEs, in particular at the lexical and syntactic level,

- enhancing the usability of MWE lexicons and valence dictionaries in parsing,
- paving the way towards interoperability of lexicons and the reduction of their production cost.

The following aspects will be studied:

- How to account for the fixed character of MWEs with respect to some linguistic phenomena on the one hand, and their similarities to regular syntactic structures on the other hand?
- In particular, how to represent, at the lexical level, phenomena most relevant to parsing, i.e., agreement, discontinuity and free word order?
- How should MWE lexicons be structured to be easily convertible and maximally reusable in different frameworks?

The expected outcomes are:

- reports on the contrastive analysis of lexical and syntactic properties of MWEs in different European languages,
- already existing lexicons and valence dictionaries enhanced and enriched with MWEs, for several European languages,
- design proposals for cost-saving abstract models of MWEs' properties, such as meta-grammars that could be automatically mapped to different lexicon and grammar formalisms; these models would apply to different languages in question.

2. **WG2** will study **PARSING TECHNIQUES FOR MWEs**.

Parsing is the fundamental process aiming at the representation of the syntactic structure of phrases and sentences. In the traditional methodology this process is based on lexicons and grammars representing roughly properties of words and interactions of words and structures in sentences. Several linguistic frameworks, such as Head-driven Phrase Structure Grammar (HPSG), Lexical Functional Grammar (LFG), Tree Adjoining Grammar (TAG), Combinatory Categorial Grammar (CCG), etc., offer different structures and combining operations for building grammar rules. These already contain mechanisms for expressing properties of MWEs, which, however, need

improvement in how they account for idiosyncrasies of MWEs on the one hand and their similarities to regular structures on the other hand. Thus, these mechanisms are rarely used on a large scale. Moreover, the semantic representation of MWEs remains a challenge. Finally, grammar development, like lexicon construction (see WG1), is considered relatively costly. Another important challenge in parsing is the linguistic ambiguity. Most phrases and sentences yield several competing parses and in some particularly complex cases this number may rise to several hundreds or even thousands. Consequently, the parsing process is slow and raw parsing results are hardly usable. Seminal works have shown that providing mechanisms dedicated to MWEs, and fully integrated into parsing, improves both parsing efficiency and linguistic precision, since MWEs enable pruning spurious parse structures.
In this context, the objectives of WG2 will be:

- a better understanding of the potential of different linguistic frameworks with respect to parsing MWEs,
- enhancing parsing efficiency,
- reducing the cost of grammar production.

The following aspects will be studied:

- Should MWEs be fully integrated in parsing or be addressed at a pre- or post-processing stage?
- How can MWEs influence parsing speed by reducing spurious ambiguity?
- How to express the semantics of MWEs in parse structures?

The expected outcomes are:

- recommendations of best practices for MWE representation and treatment in parsing within different theoretical frameworks; the resulting designs should be maximally interoperable,
- extensions of existing grammars for several European languages with rules handling MWEs,

- design proposals for abstract compact models of MWE-specific grammar rules that might be automatically mapped to different lexicon and grammar formalisms, thus reducing the production cost of particular grammars; these models would apply to different languages in question.

3. **WG3** will focus on **HYBRID PARSING OF MWEs**.

As mentioned in Section B.3, the complementarity of data-driven and knowledge-based approaches may lead to more efficient and more linguistically precise parsing tools.

On the one hand, given properly annotated corpora and treebanks, current statistical parsing methods achieve interesting results in tasks that can be modelled as sequential tagging problems. They are efficient, although there is still room for improvement, for contiguous short multi-word units, e.g. compounds, complex terms or named entities. They offer, however, a limited expressive power, which makes long-distance relations and discontinuities, typical of verbal MWEs (see WG1), hard to model. Recent pioneering works manage to model verbal expressions, although they suffer in overall parsing accuracy due to: (i) the difficulty of integrating external language resources, (ii) the scarcity of MWE-annotated training corpora (due to their high production cost), (iii) imperfect annotation schemas of the few existing annotated corpora. A substantial improvement may come from supplementing the costly annotated data by unannotated data, whose quantity is practically unlimited.

On the other hand, knowledge-based parsing methods, often capable of coping with long-distance relations and discontinuities, face the problem of linguistic ambiguity. A gain in efficiency has already been achieved by integrating statistical data extracted from corpora. These efforts should be continued and extended.

In this context, the objectives of WG3 will be:

- increasing the efficiency and accuracy of parsing methods, especially hybrid ones,
- paving the way towards using widely accessible unannotated data in order to improve grammars and models based on annotated data.

The expected outcomes are:

- recommendations of best practices of enhancing data-driven parsing with linguistic resources such as MWE lexicons and valence dictionaries, e.g. by MWE-oriented reranking of state-of-the-art parsers' results,

- recommendations of best practices of enhancing knowledge-based parsing of MWEs with probabilistic scores, in order to avoid spurious syntactic ambiguities while parsing MWEs,

- guidelines for extracting probabilistic scores from treebanks and for encoding them in lexicons (cf. WG1 and WG4).

4. **WG4** will work on **ANNOTATING MWEs IN TREEBANKS**.

Treebanks are crucial language resources whose role is to model the linguistic phenomena on the basis of real-life and wide-coverage data. They are widely used in lexicography, language learning and linguistic research. They also constitute the core of rapidly progressing data-driven methods, including statistical parsing. Few treebanks feature annotation levels explicitly dedicated to MWEs. Consequently, statistical parsers hardly ever account for the idiosyncrasies and the opaque meaning of MWEs, despite their great quantitative and qualitative importance. Thus, further linguistic applications (e.g. translation, opinion mining etc.) often fail to properly capture the contribution of these units to the structure and sense of utterances.

Nevertheless, treebanks – with or without explicit MWE-specific annotation levels – are a valuable source of information on frequencies of different linguistic units and structures, and their occurrence context. Such data can contribute to hybrid parsing methods addressed by WG3.

In this context, the main objective of WG4 will be to take a step towards enhanced MWE-aware methodologies of treebank construction, and their optimal usability in parsing.

The expected outcomes are:

- annotation guidelines for representing MWEs in constituency and dependency treebanks,

- recommendations on how to use current and future treebanks to automatically extract lexicons and probability scores addressed in other WGs.

E. ORGANISATION

E.1 Coordination and organisation

The Action will fully conform to its COST funding framework, i.e., it will function as a coordination and networking of nationally funded research activities. The representation of European countries and languages in the Action will have to be high and partners shall bring a substantial input, i.e. language resources and tools dedicated to particular languages, as explained in Section C.3. This huge amount of multilingual data and methods would be developed further within national projects. However, the existence of the Action will allow the enhancement of these resources and tools by bringing new cross-theoretical, cross-methodological and cross-lingual insights into linguistic and computational aspects in general, and into MWEs in particular. The Action will be managed by the Management Committee (MC), composed of national representatives, and reporting to the Domain Committee, according to the "Rules and Procedures for implementing COST Actions" (text reference #4154/11). The MC and each WG will meet at least once a year in different member countries. Moreover, plenary sessions will federate the efforts of WGs, in particular within workshops organised jointly with leading international conferences (e.g. European chapter of the Association for Computational Linguistics, Language Resources and Evaluation Conference, etc.). The kick-off meeting of the MC that formally starts the Action will be mainly dedicated to the efficient organisation of the WGs, to the coordination among the WGs, and to transverse actions. This first meeting will also address the election of the MC Chairman and Vice-Chairman, as well as appoint the WG Leaders, the Representative of Early-Stage Researchers (ESRs), the Coordinator of Short-Term Scientific Missions (STSMs), and the Dissemination Coordinator. These people will constitute the Action's Steering Committee (SC) responsible for efficient day-to-day management. The SC's role will be to prepare the deliberations of the MC and to closely follow the evolution of the Action. In particular the SC will fulfil the following tasks:

- control of the organisational structure of the Action, and of the fulfilment of the scientific programme,
- suggesting possible corrections needed to reach the expected milestones,
- organisation of the Action's annual workshop gathering all Action's members,
- realising dissemination by gathering and publishing the relevant documents, tools, resources, etc.,

- maintaining and updating the Action’s website,
- inviting external experts,
- preparing the Action’s annual progress reports and final report,
- ensuring the special support of ESRs within the Action,
- ensuring links between the Action and other research programmes via joint events (e.g. workshops organised within international conferences in the domain).

In order to fulfil its objectives most efficiently, the SC will meet at least four times a year, most often via a video-conference. The agenda of each meeting will be announced by the Chairman one week in advance and will be completed by suggestions submitted by other SC members.
The role of Early-Stage Researchers (ESR) is considered central to the Action. The Action recognises the ESR-specific needs addressed by the "COST Strategy towards increased support of early stage researchers" (text reference #295/09) and will encourage the implementation of the support measures put forward in this document, such as the COST family-friendly policy, the election of ESRs as WG Leaders, etc. Moreover, the ESR Representative appointed during the Action’s kick-off meeting will have the following roles:

- representing the interests of ESRs within the SC and the MC,
- promoting the Action’s Short Term Scientific Missions (STSMs) among ESRs inside and outside of the Action, and informing the candidates,
- promoting Training Schools organised by the Action,
- promoting COST’s Conference Grants offered outside of the Action.

STSMs and Training Schools will be considered one of the Action’s major instruments. They will mostly concern ESRsor PhD students. The STSM-Coordinator (appointed by the MC during its kick-off meeting) will arrange the scientific and budgetary assessment of STSM applications and reports, as described in "COST Vademecum (Part B) – Grant System".
The Action’s website will be one of the main coordination and dissemination tools. Its public part

will inform the wide public about the Action's objectives, scientific programme and organisation, as well as its latest achievements and news. It will also gather links to MWE-aware resources and tools, lists of conference and journal papers published within the Action, research events related to the Action's topic, references to Master's training programmes offered within the Action's network, etc. It will also facilitate enquiries from potential new partners interested in joining the network. In its password-protected part the website will offer management tools such as task specifications and timetables, internal reports, download/upload spaces for sharing documents, resources and tools, collaborative Wiki-type pages concerning the Action's topics, etc. The Dissemination Coordinator, appointed during the Action's kick-off meeting, will be in charge of maintenance and regular updates of the Action's website, amongst other roles related to dissemination.
In order to provide fundamental rules of project management the following milestones are foreseen:

- contrastive multilingual report on linguistic properties of MWEs,
- contrastive cross-theoretical and cross-methodological state-of-the-art report on techniques and solutions for MWE representation and processing,
- methodology of lexicon design accounting for MWEs and valence data,
- methodology of MWE annotation in corpora,
- methodology of accounting for MWEs in grammatical resources and in parsing,
- methodology of combining data-based and knowledge-based methods in MWE-aware hybrid parsing,
- meta-grammar formalism adapted to the representation of MWEs.

E.2 Working Groups

The scientific workload will be shared by 4 Working Groups (WGs) described in Section D. Each WG will be coordinated by its Leader appointed during the MC kick-off meeting. Each WG Leader will report to the MC and the SC on the progress of the scientific programme within the WG. She/he will animate and manage the WG's discussion list or forum. Some WGs may admit a further subdivision into subgroups in order to facilitate an efficient scientific exchange. For instance experts of a particular theoretical framework (HPSG, LFG, TAG, CCG, etc.) might collaborate

more closely within a subgroup of WG2. However, cross-theoretical discussions will be dominant, and as a general rule all groups or subgroups will cross language boundaries. Membership in different WGs will be open to all participants and not mutually exclusive.

E.3 Liaison and interaction with other research programmes

As mentioned in Section B.4, the Action will build on and extend current European projects: FLaReNet (ECP-2007-LANG-617001), CLARIN (212230), META-NET (ICT-NoE-249119), ATLAS (CIP-ICT-PSP-250467) and CESAR (CIP-ICT-PSP-271022). In particular, it will join the efforts of these previous projects in:

- using and extending standards for the resources dedicated to MWEs,
- paying particular attention to interoperability issues related to resources and tools.

Liaison with these previous projects, as well as with the current COST Actions MUMIA (IC1002) and IS1006, will be ensured by:

- regularly informing the respective coordinators on the Action's scientific programme and progress,
- co-organising, whenever appropriate, common events such as workshops within international conferences of the domain,
- promoting the Action's Short Term Scientific Missions and Training Schools among Early Stage Researchers included in these projects and actions.

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 Short Term Scientific Missions aimed mainly at Early Stage Researchers are clearly seen in

this Action as one of the main instruments for bringing collaboration and coherence into the Action. The special role of the ESR Representative is defined within the Action's Steering Committee (see Section E.1).

F. TIMETABLE

The duration of the Action is four years.

The timetable provided below summarises the main activities to be carried out within the Action. The initial MC meeting will start the Action, appoint the SC and discuss organisational issues. The scientific timetable for each WG will be elaborated in the first semester. Dissemination via the public website will start in the second semester. At the same time, communication and project management tools will be developed within the internal website. Both the MC and the WGs will meet at least once a year in different member countries. The same sessions will usually also comprise Action's plenary meetings and workshops in order to gather most of interested partners and reduce their travel costs. SC meetings will take place every three months, mainly via a video-conference. Each year one of these meetings will be dedicated to internal evaluation of the fulfilment of the scientific programme, and to suggesting possible corrections needed to reach the expected milestones. At the same occasion the SC will monitor the fulfilment of the evaluation plan, and revise it if necessary.

At least two Training Schools, as well as two Open Workshops (within established international conferences in the domain) will be organised. STSMs for ESRs and senior researchers are scheduled for the whole duration of the Action. Annual progress reports and the final report will be drawn up as specified in "Rules and Procedures" (text reference #4154/11).

Activity	**Year 1**				**Year 2**				**Year** 3				**Year 4**			
MC meeting	+				+				+				+			+
Scientific timetable for each WG	+															
Public website		+														
Internal website		+														
WG meeting	+				+				+				+			+
SC meeting	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Internal evaluation		+			+				+				+			
Training Schools						+								+		
Action's Workshop					+				+				+			+
Open Workshop							+								+	

STSMs for ESRs		+	+	+	+	+	+	+	+	+	+	+	+	+	+	
STSMs for senior researchers		+		+		+		+		+		+		+		
Annual report				+				+				+				+
Final report																+

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: BG, CH, CZ, DE, EE, FR, HU, IT, NO, PL, PT, RS, SE, 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 of the Action's dissemination plan include:

- the members of the Action's consortium who will closely follow the evolution of the Action, the results of all its meetings and other activities,
- Early Stage Researchers – members of the Action's consortium, of the associated research centres, and of other institutions, who will gain a better integration in the network and in the ERA through STSMs, Training Schools, and direct participation in the WGs and the MC,
- Master and PhD students, as well as candidates to life-long training, receiving training in the research centres and Working Schools organised by and associated with the Action's consortium,
- members of research projects, actions and networks dedicated to issues related to the Action's scientific programme,

- other researchers in the fields of NLP, computational linguistics, linguistics, psycholinguistics, etc.,
- professionals in the field of language industries; several representatives of European enterprises are members of the current consortium; new industrial contacts will be established in the Action's early stage,
- language resource providers,
- European institutions, dedicated to Europe's multilingual heritage,
- foreign language teachers and learners, sensible to idiomatic issues in language,
- general public interested in the use of language processing tools.

H.2 What?

A wide range of means will be used to ensure a wide dissemination of the Action's results, including:

- Action's **portal website**, which will be the main dissemination tool. It will contain: a description of the scientific issues related to MWEs dedicated to a wide public, a detailed description of the objectives, the scientific programme, the description of WGs, the list of achievements and news, links to MWE-aware resources and tools, state-of-the-art reports, training material, list of conference and journal papers published within the Action, links to research events related to the Action's topic, references to Master's and doctoral training programmes offered within the Action's network, an easily accessible contact point for potential new partners interested in joining the network, etc.
- Action's **password-protected website** – used for project management issues (see Section E), including sharing of working versions of documents and resources.
- Action's **internal mailing/discussion lists** – one for the entire consortium, and one for each WG.
- Action's activities open to a wider scientific public:

- **Training Schools** open to young researchers and to life-long training candidates. **Training material** used during these events will be published on the Action's portal.
- **STSMs** addressed mainly to ESRs,
- **Workshops** within established international conferences such as:
  - International Conference on Computational Linguistics (COLING),
  - Conference of the Association for Computational Linguistics (ACL),
  - Conference of the European Chapter of the Association for Computational Linguistics (EACL),
  - Conference on Language Resources and Evaluation (LREC),
  - Conference on Empirical Methods in Natural Language Processing (EMNLP),
  - Conference on Lexical and Computational Semantics (*SEM), etc.

- Announcements of these open events, calls for papers and calls for participation in **established mailing/discussion lists** open to a large scientific community.
- Electronically published open-access **proceedings** of open workshops organised by the Action.
- Scientific **publications** in peer-reviewed international journals and conferences in the domain of NLP.
- **Master's and doctoral trainings** in NLP carried out by the research and higher education centres of the Action's consortium.
- European infrastructures for the dissemination of language resources and tools, notably **META-SHARE.**
- **Press releases** announcing the Action's topics, objectives and achievements on the occasion of events such as workshops and plenary meetings.

H.3 How?

The use of the dissemination methods has been described in the above sections. The Dissemination Manager, appointed at the MC kick-off meeting, will be in charge of the dissemination plan management. The SC will evaluate the fulfilment of this plan on an annual basis, and revise it if necessary.