SEVENTH FRAMEWORK PROGRAMME
This document refers to PEOPLE Work Programme 2008

Grant agreement for: **Initial Training Networks**

Annex I - "Description of Work"

Project acronym: CLARA
Project full title: Common Language Resources and their Applications
Grant agreement no.: 238405
Date of approval of Annex I by the Commission: 7 July 2009

PART A:

A 1 List of beneficiaries and project summary

A.1.1 List of Beneficiaries

Beneficiary Number	Beneficiary name	Beneficiary short name	Country	Date enter project	Date exit project
1 Coordinator	UNIVERSITETET I BERGEN	UiB	Norway	month 1	month n
2	EBERHARD-KARLS-UNIVERSITÄT TÜBINGEN	UTU	Germany	month 1	month 48
3	SIA TILDE	Tilde	Latvia	month 1	month 48
4	KØBENHAVNS UNIVERSITET	UCPH	Denmark	month 1	month 48
5	UNIVERSITAT POMPEU FABRA	UPF	Spain	month 1	month 48
6	HELSINGIN YLIOPISTO	UHEL	Finland	month 1	month 48
7	UNIVERZITA KARLOVA V PRAZE	CUNI	Czech Republic	month 1	month 48
8	NORGES HANDELSHØYSKOLE	NHH	Norway	month 1	month 48
9	MAX PLANCK GESELLSCHAFT ZUR FÖRDERUNG DER WISSENSCHAFTEN E.V.	MPI	Germany	month 1	month 48

List of Associated Partners

Associated partner Number	Associated Partner name	Associated Partner short name	Country	Level of Participation (2 or 3) (*)	Organisation Status (**)
1	HUNGARIAN ACADEMY OF	HASRIL	Hungaria	2	Research organization

	SCIENCES				
2	COMPERIO AS	COMPERIO	Norway	2	SME
3	TEMIS DEUTSCHLAND GmbH	TEMIS	Germany	2	SME
4	EVALUATIONS AND LANGUAGE RESOURCES DISTRIBUTION AGENCY	ELDA	France	2	SME
5	UNIVERSITY OF ZAGREB	FFZG	Croatia	2	Higher education establishment
6	UNIFOB AS	AKSIS	Norway	2	Research organization
7	MIKRO VÆRKSTEDET	MV	Denmark	3	SME
8	INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE	INRIA	France	3	Research organization
9	LEXICAL COMPUTING Ltd	LC	United Kingdom	2	SME

(*) ***level 2****: provider of research training and complementary training including secondment or* ***level 3****: member of the supervisory board*
(**) as per Form A2.2 selection

A.1.2 Project Summary

Free Keywords:

Language sciences, Humanities, Information science, Information technology, Computational linguistics

Abstract:

CLARA will train a new generation of linguistic experts who will be able to cooperate across national boundaries on the establishment of a common language resources infrastructure and its exploitation for the construction of the next generation of language models with wide theoretical and applied significance. The scientific objectives of the CLARA research context are twofold: 1. to develop the next generation of data-intensive language models and applications by integrating approaches across language and country boundaries; 2. to contribute to the establishment of a pan-European infrastructure for language resources. CLARA will supplement basic research competencies in the language and text sciences with specialized knowledge and skills in computer science, knowledge engineering, databases, statistical processing and language and speech applications. Participation in advanced research at leading universities will be complemented with additional training at industrial partners contributing to careers in industry.

PART B:

B.1 Description of the joint Research Training Project

B.1.1 Project Overview

The principal scientific objectives of the CLARA research context are twofold:

1. to develop the next generation of data-intensive language models and applications by integrating approaches across language and country boundaries;
2. to contribute to the establishment of a pan-European infrastructure for language resources.

In this emerging new research context, training and career development of CLARA researchers will produce the next generation of linguistic scholars and engineers who master the following knowledge and skills:

- language modeling, esp. using statistical, machine learning and finite state techniques;
- language services and applications in multilingual information societies.
- multilingual technologies including contrastive studies and translation;
- encoding and annotation of language resources;
- interoperability and harmonization of language resources according to standards;
- distributed access management, large scale infrastructures and grid technology
- thorough knowledge of industrial methods and needs, to improve industrial career prospects.

In the past decades, language models have increasingly moved from manually constructed, deductive rule based systems towards systems based on statistics and machine learning of authentic data. The construction of such language models presupposes very large amounts of suitable, annotated language data for every language in question. Current approaches are often generally called "corpus-based" due to their reliance on text or speech corpora, but other, derived language resources such as lexicons, wordnets, termbanks etc. play an important role as well. Even newer approaches are aimed at building hybrid models that include both rule-based and data-based knowledge sources, e.g. weighted finite-state transducers.

CLARA has interdisciplinary relevance. The project is primarily situated in the humanities, because the use of language sources, which are increasingly digitized, is pervasive in all humanities disciplines. CLARA will also have relevance for psychological and social science approaches to language including the study of mental language processes and social dynamics of language groups. The project is also relevant beyond the Humanities and Social Sciences, since the traditional language sciences must be complemented by relevant knowledge from information theory, statistics, computer science, cognitive science and artificial intelligence, to name just a few. Thus, the next generation of language researchers will need a new combination of training components which most universities and research institutions cannot offer by themselves.

The technological applications of CLARA are intersectoral. Language models have, for instance, a huge potential in the educational sector, which currently faces the gap between advanced research in Intelligent Computer-Assisted Language Learning (ICALL) and the actual needs and practice of current Foreign Language Teaching (FLT). In the IT sector, localization, search engines and information retrieval constitute huge application areas which are in need of innovative language enabling systems. Many commercial actors in this sector are SMEs that have started up in a fairly local market situation and need to link up with other actors to take the step towards Europe-wide, cross-lingual approaches. The publishing sector, in particular publishers of printed and digital translation dictionaries, termbanks and other translation aids, is a final example of an important sector which will benefit from CLARA. Current dictionaries are not up to date, they do not cover the finer meanings of translation distinctions and they lack domain-specific terminology. Current dictionary making is also too slow to account for rapidly changing word uses. Better exploitation of language resources is marking a paradigm shift that will offer benefits in this sector.

B.1.2 Concept and Project Objectives

The principal ITN objective is the training of a new generation of linguistic experts able to cooperate across national boundaries on the establishment of a common language resources infrastructure and its exploitation for the construction of the next generation of language models with wide theoretical and applied significance. Specifically in the field of languages, there is a need to transfer theoretical models and approaches from one language to other languages, and to coordinate research on issues of multilinguality across boundaries, respecting the specific needs of different languages and language groups while striving for models that are applicable across boundaries in a tightly integrated European information society.

Recent advances in technology and widespread research efforts have expanded both the size of corpora and the extent of their annotations, e.g. in the area of deep syntactic annotation (treebanks), semantic and pragmatic annotation, and multilingual (parallel) corpora, while also various speech and multimodal corpora are becoming available. From corpora as basic resources, other resources are being derived, e.g. lexicons, frequency lists, word nets, term banks, etc. Although a large number of language resources have been produced to date, many scientific and organizational challenges remain, including the following:

- Theories and modeling approaches have not yet been applied on a wide range of languages, while some languages or language types (e.g. morphologically rich languages) may present special challenges.
- Parsers and other tools tend to be language specific (English in particular) and many tools for creating modules, resources and applications impose restrictions in their further use by SMEs and researchers.
- The gap between academic models and the needs of industrial actors who aim at real life applications remains to be bridged.
- The standardization and compatibility of language resources is still inadequate, despite the existence of metadata and integration initiatives like IMDI and DAM-LR, coding and annotation practices like XML and the TEI guidelines and semantic interoperability initiatives like ISO TC37/SC4 and LIRICS.
- There is a lack of appropriate documentation for many resources, and moreover there is no good overview of available resources for all the European languages.
- Since some resources are developed for specific purposes, there is a challenge to convert them so they can be reused for other purposes.
- There is a multitude of different conditions and restrictions for access and use in R&D.
- The long term preservation of language resources needs to be secured.
- Efficiency issues in accessing language resources in very large repositories must be addressed.

B.1.3 Scientific and technological objectives of the research and training programme

Research

CLARA has the following scientific and technological objectives in the area of language infrastructures:

- further work on standardization of coding and annotation practices and promotion of standards
- development of registries and documentation systems for language resources

- transfer and integration of single-purpose resources to interoperable, reusable and extendable forms
- development of transnational legal and organizational frameworks for simplified access
- conceptual and technical models for efficient access and preservation of language resources.

CLARA furthermore aims at achieving the following objectives:

- Transfer and extension of modeling approaches and tools to different languages and language types, through synergies between partners with complementary approaches or one partner having tools and the other data.
- New insights in commonalities and differences between languages through data-based comparative and cross-lingual studies, through synergies between partners working on different languages.
- Development and testing of systems in real life settings, through synergies between academic and industrial partners.
- Encourage researchers and SMEs to produce parsers and language processing modules which are mutually compatible for all relevant languages.

We will offer young scientists the opportunity to work in cutting-edge projects in the work packages listed and further described below. Each of these work packages focuses on one sub-theme within the overall goal of building the next generation of language models. Within each work package, complementarities between partners are actively exploited, and two secondments at industrial sites are foreseen. The joint research programme is described as follows, including a tentative breakdown of tasks among contractors.

Work package 2: Designing and Testing Common Infrastructures

Lead Participant: MPI.
Person-months: MPI 72 ESR.

Objectives:

This work package is aimed at the design of the next generation of common language resource infrastructures on which future models will be based. It will have a strong link with CLARIN to reach its goals. The following research topics can be discerned:

- sustainability and long term preservation of language resources
- construction of an interoperable language resource infrastructure
- efficient management of and access to very large language resource repositories
- organizational and technical models for large service grids
- design of a discipline-oriented service architecture
- development of new standards and promotion of existing ones

Deliverables with description of work:

Project 2A: Common Language Resources Infrastructure (MPI)

Objective: We propose to build exemplaric modern online repository with structured metadata descriptions and relevant linguistic content. Web applications and services will be used and extended to allow the manipulation and usage of resources such as annotations and lexicons via the web.

Research strategy and methodology: Relevant resources from different creators that used different linguistic encoding standards are analysed, described with metadata, transferred into archivable formats and then integrated into such an archive system. This will include the resources of the TRACTOR archive of Central and East European language resources, collected as part of a major European initiative, but currently unavailable to researchers. Applications will be written to harvest metadata from different centers to create an integrated browsable and searchable domain. Web applications will be created to create virtual collections by making use of federation middleware so that simple operations can be carried out on the virtual collection.

Feasibility, innovative aspects and relevance: The results can immediately be integrated into the CLARIN infrastructure, the technologies will conform to emerging GRID standards, and the project will make use of existing tools resources already held at MPI. The project is fully scalable in terms of the scope of resources to be included.

ESR training aspects: Training will be provided in principles of digital archiving and metadata standards, and in relevant tools.

Project 2B: Interoperability Standards and Semantic Web Technology (MPI)

Objective: We propose to tackle interoperability problems at the syntactic and semantic level. Annotation structures and lexicons will be analysed and transferred to generic schema-based formats such as LMF, XCES and EAF. The resulting resources will be integrated into repositories such that joint operations are made possible. The expected semantic interoperability problem will be tackled by integrating all linguistic concepts into a concept registry based on the ISOcat model and appropriate relations will be registered allowing cross-resource searching for example.

Research Strategy and methodology: Resources in legacy formats will be converted into generic schemas, taking into account structural and character encoding details, probable information losses and semantic interpretation aspects. The transformed resources are integrated into a repository infrastructure that allows virtual collection building. For the created virtual collections an analysis of the linguistic concepts is made. For those that match the official ISOcat definitions, references are created in a practical ontology, for those that are not contained entries are made in a project registry instance. An small RDF framework is set up to create relations between a selection of concepts so that joint search operations can be carried out exploiting the registered relations and concepts. A web application is created that allows to store and therefore re-use the elements of the created practical ontology.

Feasibility, innovative aspects and relevance: The generic formats have been tested against a wide number of structures. The ISO DCR model is stable, the ISOcat software framework has been demonstrated at ISO meetings and the API is already in use by web-applications so that there is a running concept registry framework. For creating relations tools have been developed by the semantic community which need to be extended to fit to the needs of the project. A search engine can be developed reading the resources being represented in generic formats and exploiting the ontology specifications.

Work package 3: Lexical Semantic Modeling

Lead Participant: UIB.
Person-months: UIB 36 ESR, UCPH 36 ESR, UPF 36 ESR.

Objectives:

This work package is aimed at state-of-the-art semantic modelling, consisting on the one hand of the exploration of bilingual corpora (LC and UIB) towards computer-assisted lexicography; on the other hand, word nets will in turn be used to enhance the information in corpora through advanced annotation schemes (UCPH). There will be **secondments** between UPF and TEMIS.

Deliverables with description of work:

Project 3B: Exploring bilingual corpora with Semantic Mirrors (UIB).

Objective: Lexical semantic databases and word nets have been manually constructed for a number of languages, but are still missing for several smaller languages. We propose to construct experimental word net-type lattices and thesauri that distinguish and interrelate word senses of homonymous and polysemous words) for European languages which have not yet been researched with the Semantic Mirrors method, and thus to contribute to the development of semantic applications for the languages in question.

Research strategy and methodology: Translation corpora are first lemmatized and aligned at sentence level (using e.g. TCA2) and word level. The set of possible translations of each lemma in both languages is collected, constituting the 'translational image' of the lemma. From overlap relations and correspondences among the translational images, word senses are individuated, and 'semantic fields' revealing assumed hyperonymy/hyponymy and near-synonymy relations among the senses are computed and represented as complex semilattices. Thesaurus-type databases, including further subsense individuation subject to certain set parameters, are finally derived from the semilattices. The results will be compared to results of other method of sense individuation, in particular the Sketch Engine based method (LC).

Feasibility, innovative aspects and relevance: The identification of word senses is highly relevant for applications in the area of web search, information extraction and dictionary production. The possibility of partly automatizing the development of such resources is therefore extremely desirable. Resources derived from translational corpora by means of the Semantic Mirrors method will be especially relevant for multilingual language processing.

ESR training aspects: Training will be provided in parallel corpora, alignment tools and the Semantic Mirrors method.

Project 3C: Semantic annotation of large corpora (UCPH)

Objective: Semantically annotated corpora, also for the lesser studied languages, may well prove to be a necessary prerequisite for moving into the next era of language technology where more sophisticated semantic web-based tools for i.e. semantic analysis and data mining can be developed. We will investigate (i) to which degree current wordnets (currently available for around 60 languages) are suitable as lexical semantic resources for automatic annotation of large, varied corpora (written, spoken and visual) and (ii) to which degree resources and tools for corpus annotation are transferable between languages.

Research strategy and methodology: As a necessary prerequisite for supervised annotation systems, the suggested methodology includes manual semantic tagging of training corpora for the relevant languages to be studied, including, however, investigations of whether the building of such Gold Standards can be supported by means of parallel corpora as envisaged in the Semantic Mirror or Sketch Engine methodologies (see other projects in this Work package). The Gold Standards will function as training corpora for machine learning systems. Based on Senseval results, which have now been achieved for series of different languages, as well as on the results of Semantic Mirrors methodology which has a cross-lingual, corpus-based focus, the research project will include the study of cross- and multilingual aspects of semantic corpus annotation and the possibility of technological transfer between languages.

Feasibility, innovative aspects and relevance: Semantic corpus annotation is moving into a new era where also lesser studied languages can provide the necessary background language resources required for the task. Semantic corpus annotation can provide the much wanted synergy between large text repositories and lexical-semantic resources and thereby make text collections more accessible for different kinds of humanities researchers.

ESR profile and training aspects: Training will be provided in semantic annotation and word net technology.

Project 3D: Automatic lexical acquisition (UPF)

Objectives: A large number of language models and applications (e.g. machine translation, question answering, etc.) are facing a bottleneck, as they need large lexical resources which are costly to build and maintain for every language and for every new domain. The goal of our research is to provide the automatic acquisition of lexical information to be included in computational lexica, terminology, thesauri, etc. Systems that reduce the cost of producing lexical resources will contribute to reduce the distances among language communities. Our research will work on methods and techniques that ensure portability in multilingual models and applications.

Research strategy and methodology: The performance of different techniques (including Bayesian classifiers, Decision Trees, Hidden Markov Models, Support Vector Machines) will be compared to study the class of learners that better fit the characteristics of linguistic data. The work will adhere to proposed standards for lexical information encoding such as LIRICS Lexical Markup Framework (now under ISO/TC 37/SC 4) and Data Category Registry while paying prioritary attention to lexical resources used by CLARA industrial partners.

Feasibility, innovative aspects and relevance: Current approaches do not immediately handle two crucial problems we should critically address: lexical ambiguity and missing information, which are intrinsic characteristics of natural language. These will be addressed in this project.

ESR profile and training aspects: A candidate with a background including statistics and machine learning will be introduced to lexicology, or vice versa.

Secondment to 3D: Large-scale Information Extraction (TEMIS)

In order to benefit from methods and tools available at the industrial partner TEMIS, the ESR in B4 will spend a short period at TEMIS and obtain training related to the following subproject:

Objective: The application of information extraction approaches on real-world tasks emphasize a number of parameters, that are often different from more research-oriented contexts. Ease of deployment, resource efficiency or the level of maturity for instance can all turn out as descives factors and can influence descisions more than small differences in accuracy. A focus on the assessment of these factors and their consequenences for practical information extraction is of high relevance for an industrial approach. Parameters that enter such considerations may be the availability of a training corpus, the specificity of the information which needs to be processed, the desired level of accuracy, time schedule, and many others.

Research startegy and methodology: In this programme selected information extraction tasks will be considered and investigations will explore various alternatives for their implementation that take into account the above-mentioend parameters. We will aim at making steps towards a methodology that allows to classify a task together with its side conditions (see above) according to which appears to be the best approach.

Feasibility, innovative aspects and relevance: The activities will be based on an existing Information Extraction System called Luxid, which gives access to a range of NLP methods such as Part-of-Speech-Tagging, Shallow Parsing or Sequence Labelling with Conditional Random Fields and offers interfaces for the integration of additional ones. Since the system is employed in production environments at large organisations it represents a reasonable testbed for the investigation of parameters that influence the design choices for the implementation of a given information extraction task.

ESR profile and training aspects: Reseachers will be involved in real-world information extraction tasks and the process of deciding on the proper way to address them.

Work package 4: Next Generation Domain Modeling

Lead Participant: NHH
Person-months: NHH 30 ESR, TILDE 30 ESR

Objectives:

This work package is aimed at the harmonization of terminological resources across Europe.

Deliverables with description of work:

Project 4A: Harmonisation of Terminological Resources (NHH, TILDE, UIB, AKSIS)

Objective: For professional and technical language resources, the harmonisation of language resources across Europe poses problems at both conceptual and technological levels. Although there are a large number of terminological resources in Europe, those resources are fragmented, located in different institutions and in different formats. We propose to develop language-independent methods and tools for constructing and consolidating multilingual termbases and ontologies, for corpus-based term extraction and for ontology-based domain recognition of text. Moreover, we propose to investigate the degree of compatibility of different term bases at the top level of domain classification, at the mid-level of sub-domain, and at the lowest conceptual level, i.e. the concept itself and its hierarchical relations to neighbouring concepts within the sub-domain. The project aims to develop new, innovative approaches to cross-termbase integration.

Research strategy and methodology: The training will be organized at NHH (with the cooperating partners UIB and AKSIS) and TILDE in order to offer ESRs training on different aspects of terminology work and paying special attention on semantic enrichment and integration of terminological resources. Domain-specific resources are used iteratively to identify sections of the general corpus with a high content of text from the given domain, to provide an initial classification of the general corpus. New domain specific term candidates are extracted semi-automatically from the machine-classified texts, which are in turn fed into the pool of domain-specific resources after manual check. Finally, term candidates undergo a concept-based match with terminological resources representing other languages and are integrated into a common multilingual termbase which is compiled in accordance with accepted metadata standards (IMDI).

Feasibility, innovative aspects and relevance: The proposed methods, tools and domain-specific terminologies are beneficial to end-user systems like machine translation, text summarisation, text generation etc. Automatic domain recognition of text may also improve the precision of taggers and parsers, word sense disambiguation, etc. The online terminology data bank facilitates terminology data accessibility and exchange, standards of EuroTermBank project facilitates terminology interoperability, sharing and reuse. Finally, multilingual term bases are crucial for efficient (manual) translation and instrumental to the preservation of European cultural-linguistic heritage and as a preventive against domain-loss in smaller languages.

ESR training aspects: Training will be provided in term base technology, term extraction methods, term classification approaches, ontology construction, metadata and annotation schemes, industrial best practice, and will include **secondments**.

Secondment to 4A: Investigation of (semi) automatic import strategies for existing thesauri and terminologies (TEMIS)

In order to benefit from additional training in methods and tools at the industrial partner TEMIS, ESRs will be detached at TEMIS for the following subproject.

Objective: A critical issue for the efficient deployment of methods of information extraction/automatic indexing (IE) is the rapid definition of analysis grammars and rule sets. Ideally existing thesauri can be used, but these are often not directly useable for automatic processing. Using appropriate heuristic and linguistic methods that take into account the structure of both the existing terminologies as well as the target representation of the NLP grammar is decisive to make this process manageable.

Research Strategy and Methodology: In this subproject the task to define and apply appropriate linguistic and heuristic methods to existing thesauri and terminologies will be studied using a finite-state information extraction platform as a reference.

Feasability and innovative aspects and relevance: The activities will be based on an existing Information Extraction System called Luxid, designed by TEMIS.

ESR training aspects: Reseachers will be involved in real-world information extraction tasks and the process of deciding on the proper way to address them.

Work package 5: Multimedia and Multimodal Communication Modeling

Lead Participant: UCPH
Person-months: UCPH 36 ESR, MPI 24 ER

Objectives:

The study of human verbal behavior has long been isolated from nonverbal ways of expression even if they are inherently linked. In order to study human communicative behavior empirically, we need to build multimedia and multimodal resources, requiring new annotation schemes to be designed and applied. The projects below complement each other. The UCPH multimodal corpus and the connected annotation scheme for non-verbal behaviour can be used as testbed for the MPI research, while the results at MPI will verify if the UCPH categories are useful to classify regular patterns.

Deliverables with description of work:

Project 5A: Multimedia and Multimodality Resources and Technology (MPI)

Objective: We propose to build smart pattern recognition components and integrate them into a widely-used multimedia annotation framework so that new layers of annotations on sound and/or video recordings are created which can then be exploited by existing flexible search frameworks. The modules can be of overseeable complexity such as a silence/signal detector or more complex by cascading.

Research strategy and methodology: Typical audio and video recordings are studied to detect regularities and to then allow the participants making a formal description of the patterns over time. These formal descriptions are transferred into a software component that can be integrated into the existing ELAN framework and that creates an additional layer of annotation. It will be investigated for each component whether the existing flexible search engine needs to be extended to allow the combination of the created annotations with the existing ones to yield the expected recognition results. In particular we foresee the extension towards stochastic and learning methods.

Feasibility, innovative aspects and relevance: The paradigm has been investigated already by senior researchers and essential steps for a flexible extendible framework have been made. Different modules can be combined and their effects and combination with other patterns will lead to new insights and will inspire theorizing which will be an important step to overcome the limitations of purely stochastic systems such as Hidden Markov Models or Artificial Neural Networks.

Project 5B: Empirical Studies of Multimodal Behaviour (UCPH)

Objective: One line of research in multimodal behaviour individuates non-verbal patterns such as the detection of eye, hand and body movements and trains algorithms to interpret these movements.

Another line of research aims at recognizing the communicative functions of non verbal behaviours and studying the interaction of these behaviours by developing multimodal corpora, creating annotation schemes and building up reliably annotated multimodal corpora covering different cultural and social settings. The present project has the goal to produce insight on essential aspects of multimodal behaviour such as feedback, dialogue management, information structure and reference by conducting empirical studies in audiovisual corpora based on manual annotation and analysis as well as automatic methods such as machine learning.

Methodology: Existing multimodal corpora as well as multimodal data annotated within the project will be used to produce empirically-based evidence of the way in which specific non-verbal behaviors are used in human communication. Relevant research issues are for example to which extent this behaviour may be captured in formal categories that an algorithm can be trained to recognize, how crucial the cultural context is to interpret the behavior, or whether a universal core can be defined for non-verbal behaviours, similarly to what certain linguistic schools have attempted to do for language. Analyses of multimodal communication in different cultural settings, communication situations and social contexts are also relevant.

Work package 6: Applications

Lead Participant: UTU.
Person-months: UTU 24 ESR, TILDE 42 ESR, CUNI 36 ESR.

Objectives:

This work package is concerned with ICT applications of next generation language models in key sectors: the educational sector and the language services sector (translation, proofing, information retrieval).

Deliverables with description of work:

Project 6A: Connecting ICALL Research and Development to Foreign Language Teaching Needs (UTU)

Objective: The project will investigate and bridge the gap between research in Intelligent Computer-Assisted Language Learning (ICALL) and the actual needs and practice of current Foreign Language Teaching (FLT) – a crucial step towards the next generation of intelligent FLT tools.

Research strategy and methodology: The gap between ICALL research and FLT practice is explored by (a) identifying needs and shortcomings in FLT practice, (b) reviewing the results from Second Language Acquisition research, in particular concerning the role of noticing and awareness of language forms and of feedback, (c) designing activity types which can fill those needs, (d) designing the language model, activity model, and learner model, (e) developing and reusing NLP approaches and resources, such as those developed in the network, to analyze and provide the individualized feedback to the learner, and (f) evaluating the resulting ICALL activities with real-life learners. The analysis of learner input in ICALL systems is also closely related to the annotation of learner corpora (cf., e.g., the recent CALICO workshop on the topic), for which the annotation tools and resources developed and made available in the network are directly relevant.

Feasibility, innovative aspects and relevance: The research and training components of this project can rely on a concrete starting point, the TAGARELA system, which Detmar Meurers and Luiz Amaral developed and successfully used in Portuguese courses at the Ohio State University. On this basis, the project can successfully investigate which activity types needed in FLT practice can be integrated, where the necessary NLP analysis is available and effective, and how explicit learner modeling can improve the disambiguation of the learner input.

Project 6B: Proofing tools and resources for under-resourced languages (TILDE).

Objective: We propose knowledge transfer to develop proofing tools (mainly grammar checkers) for languages that currently do not have these technologies available.

Research strategy and methodology: Grammar checkers for world's major languages are developed and included in text processing software. However majority of small languages luck this feature. Grammar checkers are not much researched during the last decade; they typically use shallow syntactic structure rules to find agreement and word order errors and stylistic problems. This approach does not give good results for highly inflected languages with rather free word order. The first research direction will focus on multilinguality of grammar checking techniques. Existing grammar checking techniques will be evaluated to find language specific phenomena and new tools will be developed. The second research direction will be an investigation of the existing grammar checking methods and development of new methods to create suitable techniques to process deep syntactic structures of highly inflected languages.

Feasibility, innovative aspects and relevance: Good proofing tools are important for any language to survive and develop, but they are crucial for small and under resourced languages. Grammar checkers for highly inflected languages are not well researched and developed. Europe is multilingual with many languages; research groups working on proofing tools for under resourced languages are quite isolated and scattered through Europe therefore a network for gathered research and knowledge exchange is necessary.

ESR training aspects: Training will be provided in parsing technologies, text processing, CFG and dependency grammars.

Project 6C: Translation tools and resources for under-resourced languages (TILDE, FFZG and CUNI).

Objective: Despite significant progress in machine translation (MT) technologies globally, the circle of languages and language pairs that take advantage of these technologies remains limited. These technologies are not available for a number of languages in Europe, for others they are often restricted to only one language pair – the national language and English. We propose to investigate possibilities to facilitate development of translation tools and resources for languages that currently do not have or have limited translation technologies and resources.

Research strategy and methodology: The first research direction will be an investigation of the existing MT methods to find the most suitable solution (rule-based, data-driven, hybrid) for highly inflected languages with limited or no parallel corpora available. The second direction will be research of innovative content creation methods (mainly parallel corpora, and lexicons) to find effective methods to create content necessary for data-driven MT technologies. The third direction will be the base language technologies and resources (lemmatizer, taggers, parsers and dictionaries) for under resourced and highly inflected languages. CUNI will provide large-scale resources for Czech, an inflective language targeted in this project, and will use a statistical paradigm based on phrase- and tree-based transfer models with linguistic features. This project will be ***lab-based*** as CUNI will use time on its High-Performance Computing cluster.

Feasibility, innovative aspects and relevance: The current research in MT field is mostly English centered and assumes availability of large parallel corpora. We propose to do MT research for highly inflected languages with rather free word order and limited parallel corpora available. Research groups working on MT for under resourced languages are quite isolated and scattered through Europe therefore a network for gathered research and knowledge exchange is necessary.

ESR training aspects: Training will be provided in different translation technologies, parallel corpora, text processing, lexicon building and alignment tools.

Work package 7: Parsing Technologies and Grammar Models

Lead Participant: CUNI.
Person-months: CUNI 36, UHEL 48, UTU 48, UIB 36.

Objectives:

This work package is aimed at the exploration of next generation parsers. New finite state approaches will be developed for the construction of hybrid models. Furthermore, the next generation of analyzers must perform a deeper analysis in order to support meaning extraction. This requires a careful tuning of analyzers to empirical data in treebanks, i.e. syntactically annotated corpora, and research into deep structural commonalities and differences between languages.

Deliverables with description of work:

Project 7A: Finite-state parsing methods (UHEL and UTU)

Objective: Consistent processing the some 100 languages, including European official and otherwise relevant languages is necessary for achieving common European platforms. This requires adequate computational metods and software, solid linguistic principles and common standards. Methods using finite-state transducers (FSTs) have proven to be a general framework for describing morphological and phrase level processing of any European language and provide a technically uniform platform for the implementations. New possibilities for relating the FST algorithms to language models and formalisms for parsers have emerged due to recent innovations in the methods for compilation. The aim of the project is to deepen the understanding of these methods and invent new applications for large scale language processing tasks such as shallow parsing and information extraction.

Research strategy and methodology: One of the aims is to combine rule-based models and probabilistic methods through the use of weighted FSTs. Formal elegance and practical efficiency can be pursued by the composition of various language models once they are represented as weighted or unweighted FSTs. New methods allow for the compilation of a wide variety of constraints. The Generalized Restriction (GR) operator enables the implementation of sophisticated rule-based constraints, and the Optimality Operator allows for the implementation of discrete-valued probabilistic constraints. Both of these rule-based approaches can be combined with traditional statistically based methods, which provide the weights for other parts of models. This mixture of rule-based and probabilistic approaches will provide a powerful and elegant approach to practical problems in NLP.

Feasibility, innovative aspects and relevance: The methods to be created are new, better adjusted for modelling of different aspects of language than those of Xerox XFST or Helmut Schmid's SFST. Still, FST techology is well understood and feasible. The innovations enable an almost infinite range of language models to be expressed in a technically identical form as weighted FSTs which can be combined (and composed) in unrestricted ways. The methods and the resulting FSTs are language independent. Results are readily applicable for SMEs for building commercial modules and researchers for building open source modules.

Project 7B: Next Generation Deep Grammar Models (UIB and AKSIS)

Objective: Commonalities and differences between structural properties in languages have been studied, but not yet by means of a large scale multilingual corpus analysis. We want to determine to what extent the development of parallel deep grammars for typologically diverse languages may support the automatic derivation of high-quality parallel treebanks for the languages, suitable as a basis for a deeper theoretical understanding of the ways in which syntactic functions, semantic roles and translation are interrelated.

Research strategy and methodology: We propose to construct aligned multilingual treebanks as parsed corpora for a number of languages. The corpora are batch parsed by means of the XLE (Xerox Linguistic Environment) and disambiguated by means of the LFG Parsebanker. This project will be ***lab-based*** as it will use time on a High-Performance Computing cluster at Unifob.

Feasibility, innovative aspects and relevance: Parallel treebanks aligned at phrase level are innovative resources for gaining new insights in translational correspondences at structural levels. The project depends on the availability of LFG grammars. A large LFG grammar has been developed for Norwegian in the NorGram project led by Prof. Helge Dyvik, while other LFG grammars are being developed for several other languages in the ParGram project.

ESR training aspects: Training will be provided in large scale grammar development, treebanking, and use of the XLE parser and LFG Parsebanker tools.

Project 7C: Linguistic Analysis for Treebank Annotation (CUNI and UTU)

Objective: We propose here to specify a common core of syntactic and semantic features for a range of languages (Czech, English, German, Slovak, and others). The development of a common core of such categories and features will greatly facilitate the comparability of treebanks for different languages and overcome the current state of the art, where the lack of comparability of such resources has been recognized as a serious problem for theoretical and computational research. We will test and validate the resulting annotation schemes by performing annotation of texts for these languages, providing as the final product annotated corpora (treebanks) for both linguistic research and statistical language learning (i.e., creating tools for automatic syntactic and semantic analysis).

Research strategy and methodology: A corpus is morphologically analyzed and pre-parsed by state-of-the-art disambiguation and parsing tools. At CUNI, the intermediate data is then manually disambiguated by the annotators (linguists, mainly) using the TrEd graphical UI. At UTU, the Annotate tool is used to semi-automatically select the correct analysis for a given sentence.

Feasibility, innovative aspects and relevance: A highly relevant research issue concerns the synthesis between dependency-based and constituency-based annotations. Bringing together two prominent institutes in the two respective traditions will provide a good basis for an innovative solution to this important desideratum. The project is imminently feasible given that the manual and automatic tools are already available or can be easily adapted. At CUNI, the PML is easily adaptable for novel linguistic phenomena to be specified and annotated by the ESRs under the supervision of the CL professors at CUNI. Then, new tools can be developed or existing tools improved by machine learning based on the annotated corpora. At both CUNI and UTU, language independent tools for annotation and error detection can easily be applied to new languages and the adapted annotation schemes to be developed.

List and schedule of milestones

Milestone no.	**Milestone name**	**WPs no's.**	**Lead beneficiary**	**Delivery date from Annex I**	***Comments***
1	TERMCOURSE	8	NHH	9	Jun 2010, Bergen, 35 ERD
2	WORDNETCOURSE	8	UIB	9	Jun 2010, Bergen, 10 ERD
3	LRSCHOOL	8	MPI	10	July 2010, Nijmegen, 10 ERD
4	MORPHCOURSE	8	UHEL	12	Sep 2010, Budapest, 10 ERD

5	EVALCOURSE	8	UCPH	12	Sep. 2010, Paris, 20 ERD
7	TREECOURSE	8	UTU	15	Dec. 2010, Prague, 16 ERD
8	INFRASCHOOL	8	MPI	22	July 2011, Nijmegen, 10 ERD
9	ANNOTSCHOOL	8	UCPH	23	Aug. 2011, Copenhagen, 30 ERD
10	CAREERSCHOOL	8	UIB	24	Sep. 2011, Dubrovnik, 12 ERD
11	NEWDEVSCHOOL	8	CUNI	27	Dec. 2011, Prague, 20 ERD

List of Deliverables – to be submitted for review to EC

Del. no.	**Deliverable name**	**WP**	**Lead beneficiary**	***Estimated indicative person-months***	**Nature**	**Dissemination level**	**Delivery date (month)**
1	REPORT1	1	UIB	0	R	PU	12
3	REPORT2	1	UIB	0	R	PU	24
3	REPORT3	1	UIB	0	R	PU	36
4	JOINT TRAINING PROGRAMME	8	UIB	0	O	PU	36
5	REPORT4	1	UIB	0	R	PU	48
6	PROJECT2A	2	MPI	36 (MPI)	O	PU	48
7	PROJECT2B	2	MPI	36	O	PU	48
9	PROJECT3B	3	UIB	36	O	PU	48
10	PROJECT3C	3	UCPH	36	O	PU	48
11	PROJECT3D	3	UPF	36	O	PU	48
12	PROJECT4A	4	NHH	30 (NHH) + 30 (TILDE)	O	PU	48
13	PROJECT5A	5	MPI	24	O	PU	48
14	PROJECT5B	5	UCPH	36	O	PU	48
15	PROJECT6A	6	UTU	24 (UTU)	O	PU	48
16	PROJECT6B	6	TILDE	21	O	PU	48
17	PROJECT6C	6	TILDE	21 (TILDE) + 36 (CUNI)	O	PU	48

18	PROJECT7A	7	UHEL	48 (UHEL) + 24 (UTU)	O	PU	48
19	PROJECT7B	7	UIB	36	O	PU	48
20	PROJECT7C	7	CUNI	36 (CUNI) + 24 (UTU)	O	PU	48
TOTAL				570			

Training

	Early-stage and experienced researchers to be financed by the grant agreement			
Network Team	Early-stage researchers (ESR) (person-months) (A)	Experienced researchers (ER) (person-months) (B)	Visiting Scientists (VS) (person-months) (C)	Total (A+B+C)
UIB	72	0	0	72
UTU	72	0	3	75
TILDE	72	0	0	72
UCPH	72	0	2	74
UPF	36	0	0	36
UHEL	24	24	0	48
CUNI	72	0	0	72
NHH	30	0	3	33
MPI	72	24	0	84
Total	522	48	8	566

The network as a whole undertakes to provide a minimum of 582 person-months of Early Stage and Experienced Researchers whose appointment will be financed by the contract. Quantitative progress on this, with reference to the table contained in Part C and in conformance with relevant contractual provisions, will be regularly monitored at the consortium level.

Principal training objectives and methods

To achieve our goals, ESRs and ERs require an excellent understanding of basic research concepts, methods and tools in an emerging interdisciplinary and intersectoral field comprising both linguistics and the areas of information theory, computer science and statistics. In addition, ESRs and ERs should be able to work independently with at least one specific research method and have the skills to use at least one specific research tool, eventually without supervision. By the end of the project, all fellows should have a very good level of understanding of the field as a whole, being able to work with language resources in general, exploring new methods under supervision, and being able to teach the theory on which the technique is based at the level of a graduate course.

We offer the ESRs and ERs:

- Local training courses in the theory, methods and tools required for their research
- Thematic workshops that will focus on crucial issues and innovation aspects in workpackages
- CLARA-wide courses where language processing tools and applications will be demonstrated
- Onsite training under personal guidance by a scientific supervisor
- Organisational skills by involvement in organisation of network activities

- Supplementary training in career-building skills such as industrial planning and management.

We offer the following to the wider research community, partially through the CLARIN link:

- Technical and thematic workshops open for a limited number of young researchers outside CLARA
- Organisation of a summer school where advances in common language resources and their applications, including new findings in CLARA and CLARIN, will be taught to young scientists
- Organisation of thematic sessions at CLARIN meetings
- Presentation of CLARA conclusions at an EACL or LREC conference.

Rationale of the training programme

One of the principal aims of the network — to educate a group of young scientists in **state-of-the-art** approaches — will be reached by the opportunity that CLARA brings together top research teams in Europe that are willing to share language resources, tools and expertise in a spirit of intensive collaboration and exchange. Effective coordination of the workpackages requires a knowledge of each other's methods and advances. Therefore CLARA-wide activities will be organized, including courses as well as coordination activities for the supervisors. The outreach to a broad field of young scientists will take form as summer schools where ITN scientists and invited speakers will teach an overview of the state of the art in common language resources research and their applications. Furthermore, ESRs and ERs will have the opportunity to follow courses as required by individual university graduate programmes. To each fellow, a personal supervisor will be assigned.

To widen **career prospects**, ESRs and ERs are expected (where relevant) to participate in workshops at their host institute and the summer school, and to present papers at international research conferences, preferably together with other members of their research teams, and to participate in research project application procedures. While it is expected that each fellow will have a good command of English on appointment, the development of expressive language skills will be an important training consideration. Written communication skills are critically important and will be developed through tutoring and practiced through the publication of research papers in international scientific journals and the distribution of internal reports throughout the network. Additional perspectives for careers in industry will be provided by a special industrial training course.

Content of the CLARA training programme

The interdisciplinary nature of the research field lends itself to intensive and varied training aimed at developing supplemental knowledge and skills which were not part of the fellows' education, but which are essential for carrying out the envisaged advanced research. It will be mandatory for ESRs, and recommended for ERs, to attend the workshop program. The topics and the timing of the workshops are related to the time when the corresponding knowledge is required by the fellows to successfully conduct their research. Therefore, the main body of ITN-wide training will take place in the first two years of the contract. Fellows will also be stimulated to take relevant regular courses at their host institutions at the start of their stay. During the second part of the ITN, emphasis is on the exchange of insights and results and hands-on training under close individual supervision, and acquisition of complementary skills through secondments and additional courses.

Core skills to be developed in the network relate to key methodologies of modern data-based approaches to language modeling, including corpus building, coding and annotation, search and exploitation, as well as statistical models and tools. Specific skills are related to specific work packages (cf. B2). Training activities will consist primarily of local in-house training, supplemented by a series of CLARA-wide activities described below.

Network-wide events such as training courses will also be open to external researchers and will therefore in addition serve a purpose in the dissemination and exploitation of the project. In particular, researchers in the CLARIN project will be targeted.

In-house training, supervision and integration; transfer of specific and general research skills

CLARA training is first and foremost on the job training and learning by doing. At the time of contracting, ESR will be assigned a *personal supervisor* (ERs will also get one on demand) and every ESR and ER will be integrated in the research group at the host institution. At the same time, ESRs and ERs will also be informed of their contractual rights and obligations and of the training opportunities that the network offers.

Through personal supervision and integration in the research group, ESRs will ERs will be given specific skills needed in their project, including training with computer systems, as well as general research skills, communication skills, will be briefed on research ethics and will be involved in project management.

ESRs, ERs and VSs will be asked to give regular presentations of their work in progress, engage in discussions and circulate drafts of articles in their respective research groups. Special attention will be paid to the development of writing and presentation skills through personal tutoring and feedback from the supervisor and research group.

Career Development Plan, local in-house training, supervision and integration

The different on-site and off-site training activities will be integrated in a Career Development Plan that is established by the ESR/ER together with her/his supervisor at the time of contracting. This plan will include the following specifics for each appointed researcher:

1. scientific objectives of the task
2. plan of personal research activities
3. supervision plan
4. training plan including courses and attendance of scientific meetings
5. plan for the acquisition of complementary skills (proposal writing, presentations, project management, research ethics, etc.)
6. personal milestones with respect to research results
7. dissemination plan including presentations and publications.

The Career Development Plan will form the framework for monitoring progress on a continual basis through *semiannual reports* for each ESR/ER. The Career Development Plan and the seminannual reports will be written and signed jointly by the ESR/ERs and their supervisors. The person-in-charge at each site, as well as the work package leader, the Network Coordinator and Training Coordinator, will receive all plans and semiannual reports on progress and may use these as the basis for potential adjustments to the training plan.

Role of the consortium partners and contribution of industrial partners

Each level 1 partner will host ESRs or ERs and some will host a VS. They will be responsible for personal supervision of the ESR/ERs and their projects.

The associated partners in CLARA, six of those being industrial partners, contribute important training aspects, including viewpoints and experiences from research institutions and industry which the academic partners cannot offer. AKSIS, a research company, cooperates closely with UiB in two WPs and will participate in supervision, while TEMIS will host secondments (see description of WP 4). HASRIL and LC contribute content and tools to thematic training courses (as described in WP 8) and ELDA contributes to a thematic workshop (see WP8). Furthermore, COMPERIO and FFZG will, in cooperation with UIB, offer an Industrial Career Training Course where they contribute with concrete R&D experiences in a commercial setting as well as enterpreneurial and managerial expertise (see WP8).

Additional training and skill enhancement by industrial partners

The improvement of **prospects in industrial careers** will depend on being in touch with industrial needs and methods. This is in part achieved by secondments at industrial partners and by their participation in courses and supervision, as indicated above. Furthermore, a special complementary skills course offering additional training is offered to all researchers in the network. *Attendance at this course will be* ***mandatory*** *for all appointed fellows*. This training course is aimed at providing the next generation of scientists with the complementary skills that are necessary to move from theory to a marketable products and solutions. The course will include planning, market analysis, entrepreneurship, exploitation of research results, project management, proposal writing, communication, research ethics and IPR management. The course will both demonstrate a complex industrial system based on multiple components and give an introduction to strategic product planning covering the span from idea to deployment. FFZG will present the case study of the system for semantic analysis of newswire texts that is being developed for Croatian News Agency (HINA). This case study demonstrates how a complex system can be build from simple existing modules such as lemmatization, POS/MSD tagging, named entity recognition and classification, document classification, keyword extraction; with perspective to widen the customer support with more advanced Knowledge Technology such as social network analysis, event detection, trend detection etc. The second part of the course starts with an overview of the principles of market and competitor analysis, followed by a session focused on project management. Basic concepts and benefits of agile development strategy (*Scrum* methodology) will be demonstrated by examples from industrial projects. The strategy of customer-driven innovation will be explained in detail, and methods will be taught to package experience and Best Practices in product development. The workshop will also discuss a selection of architectural issues, such as best practices in the implementation of clean, backward-compatible and standard-compliant programming interfaces (APIs); as well as the design of service-oriented architectures tailored to the next generation of information access technology solutions. Further modules will include complementary skills such as proposal writing, management, communication skills and ethics.

CLARA Thematic Training Courses

A number of thematic training courses will be offered throughout the network, with the **participation of associated (including industrial) partners**. These are intended to deepen knowledge on theory and methodology that is directly needed for planned research activities. Priority is given to themes that span across several sites in the network. Each training course will typically span three full days. Most of these are planned during the initial phase, i.e. the first half of the ITN period (cf. B2 for details on these courses).

CLARA Summer and Winter Schools

Clara will offer Summer and Winter Schools which cover broad current topics and will feature internationally prominent lecturers. These schools are one or two weeks long and will also be open to outside attendees, in particular to selected, eligible researchers from the 128 CLARIN members from 32 different countries, which will contribute to wide dissemination from the action (cf. B2 for details on these schools).

TERMCOURSE: Thematic Training Course on Methods and Technologies for Consolidating and Harmonising Terminological Resources (NHH, TILDE, UIB and AKSIS)

Goal: The workshop is aimed at providing the ESR/ERs the necessary skills to utilize existing language resources in new, innovative ways, for the overall purpose of harmonizing Europe's terminological resources. The main aim is to ensure the transfer of knowledge from staff within and beyond the CLARA network working on relevant R&D projects, in order to enable participants to specify aspects on which to focus their own research efforts.

Focus: The focus will be on theoretical and technical aspects, and ESRs/ERs will be given advanced training in topics relevant for pan-European integration of terminological resources. Theoretical aspects include a) the organization of knowledge and its representation in structured databases, b) classification of domains and sub-domains, c) methods for investigating domain classification across term bases. Technological aspects include a) knowledge of existing term bases and their formats, b) knowledge of existing terminological standards and metadata schemes; c) corpus-based term extraction technology; d) multilingual integration and interoperability.

Participants: Given the interdisciplinary nature of the workshop, the course will be relevant for PhD scholars and/or postdocs who have a basic training in a relevant theoretical or technical field (terminology, linguistics, knowledge management, computational linguistics, termbase technology, language resources) but who need to increase their knowledge in relation to computational methods in terminology work.

Procedure: The course will be given as a combination of lectures, group tutorials and hands-on training in the use of resources and technology.

MORPHCOURSE: Thematic Training Course on Processing Morphologically Rich Languages (UHEL and HASRIL)

Goal: Morphologically-rich languages like Turkish, Finnish, Hungarian, etc., present significant challenges for natural language processing applications due to their relatively free word order and highly productive morphological processes (inflection, agglutination, compounding). The course will work on problems due to dictionary size, sparse data, poor language model probability estimation, high out-of-vocabulary rate and information gaps on related lexical items.

Focus: The course will introduce advanced modelling techniques addressing these problems, such as decomposition of complex word forms into smaller units, relating inflectional variants to root forms (lemmatization), methods for optimizing the selection of units at different levels of processing, novel probability estimation techniques, and the creation of a new class of data resources and annotation tools. Newer finite state techniques have proved to be useful and can be optimized in combination with other methods. The course will conclude with an assessment of present day standard techniques and a demonstration of practical applications focusing primarily on Hungarian and Finnish.

WORDNETCOURSE: Thematic Training Course: Ontologies and Wordnets and Their Use in NLP Technologies (HASRIL, UIB, LC and UCPH)

Goal: In recent years both wordnets and more formal ontologies have grown to be crucial background databases for various applications. WordNets are being used in word sence disambiguation, machine translation, information extraction and information retrieval, just to list some application areas. Over 60 wordnets have been developed over the world. Languages that are typologically different than the main model language, English, had to face additional linguistic tasks when constructing their semantic network. With theoretical focus on representing verbal structures in a lexical semantic network, the Hungarian WordNet was the first one to deal with lexicalised event structures in a systematic way. For some languages, wordnets do net yet exist and new compuational methods (such as Semantic Mirrors) are being explored to support their creation.

Focus: The course will, make a clear distinction between formal and linguistic ontologies, and give a theoretical overview of general questions concerning WordNet-building, such as a comparison of different methods (fully automated and semi-automated methods). Problems of languages that may express event structure through lexical means, e.g. through prefixes sensitive to aspect and Aktionsart, will be highlighted, and NLP applications relying on WordNets in general as well as the verbal WordNet, will be discussed in order to give a taste of practical issues. Finally, methods for automatically creating and exploring lexical relations, including SketchEngine and Semantic Mirrors, will be taught.

TREECOURSE: Thematic Training Course on Methods and Technologies for Consolidating and Harmonising Treebank Annotation (CUNI and UTU)

Goal: The course focuses on two aspects: the resources and tools as well as the linguistic analysis underlying the annotation schemes of treebanks. It will support the exploration and comparison of dependency-based and constituency-based annotation schemes and will provide the ESRs with the necessary skills to utilize existing resources and tools for treebank annotation and to adapt the them to the revised and harmonized annotation schemes.

Focus: The focus is on the comparability of existing annotation schemes for treebanks across a number of languages. It adresses both the analytic basis and the prospects for harmonizing annotations across different languages and across different linguistic frameworks.

Participants: The event will bring together the ESRs from CUNI and UTU, but will be open to all interested ESRs, especially those dealing with annotation of corpora, and researchers from other CLARA training sites.

Procedure: The course will be given as a combination of lectures, group tutorials and hands-on training in the use of resources and technology.

EVALCOURSE: Thematic Training Course on Evaluation of Human Language Technologies (ELDA and UCPH)

Goal: For any HLT research effort to be successful, it is essential that it be assessed through rigorous evaluations of the developed technologies. This allows performance benchmarking and a better understanding of possible limitations and challenging conditions. The workshop aims at providing ESRs/ERs the background and skills to use and implement state-of-the–art evaluation tools and techniques for speech technologies, grammars and parsing, machine translation and speech-to-speech translation, information retrieval/filtering, multimodal interfaces, etc.).

Focus: The workshop will elaborate on the role of evaluation on the research progress, on the need for a truly European infrastructure for HLT evaluation, on the main reasons to promote an international dimension of the evaluation, insisting on the multilingual issues. The workshop will introduce and describe some evaluation concepts (comparative evaluation versus competition, technology evaluation versus usage/usability evaluation). It will also describe the different types of evaluation and how to ensure that evaluation does not kill innovative not-yet-mature approaches.

Participants: Given the interdisciplinary nature of the workshop, the workshop will be relevant for any ESRs/ERs who are involved in the development of algorithms and systems for speech technologies, machine translation, parsers, information retrieval, multimodal interfaces, etc..

Procedure: The workshop will be given as a combination of presentations, group tutorials and hands-on training in the use of evaluation technologies for HLT.

LRSCHOOL: CLARA Summer School in Advanced Resource Creation, Archiving and Usage (MPI)

Goal: Young researchers will be trained in how to use modern technology to create language resources in particular when the source material are multimedia streams, how the resulting complex resource types can be archived, how they can be accessed via state-of-the-art web applications and how they can be enriched. It will also be shown how virtual collections can be built and how operations can be carried out on such collections. The result must be that young people have a deep understanding about modern methodologies and technologies to create, archive and use sharable resources.

Focus: The focus is (1) on using state-of-the-art tools to create resources that adhere to open standards such as XML and MPEG, (2) on teaching of how to optimally make use of open archives, how to use converters for various data types and how to define the necessary access permissions, (3) on existing frameworks that allow accessing the archived resources via various ways (from metadata up to web applications for complex objects), (4) on existing methods of how to create and use virtual collections and (5) on existing frameworks to make comments and draw relations between resources and resource fragments.

Participants: We expect young PhDs and/or Postdocs who are going to work on/with language resources in their work and who need to be educated to use state-of-the-art tools. We expect some knowledge about computational aspects, but will not require deep skills about XML schema or software programming. Thus we address those who see themselves as being users of modern technology and methodologies.

Procedure: The participants bring own resources with them such that all indicated steps can be carried out using and combining their material. The LAT technology from MPI will be used during the course. Some technology form others will be used to do efficient processing of specific tasks (speech analysis, conversion, etc).

Teachers: Mainly members of MPI will give the courses. In addition we will invite specialists who have deep knowledge about relevant standards in our domain, about audio/video codecs and appropriate software, about speech analysis and appropriate software, about semantic web techniques and representation standards such as RDF. The MPI experts have 8 years of experience in giving such courses twice per year to the indicated group of people.

INFRASCHOOL: CLARA Summer School on Infrastructure Tool Development (MPI)

Goal: Young researchers will be trained in an advanced course about all aspects that are relevant for making use of the emerging CLARIN infrastructure and about how to actively contribute to it with new resources and tools. The result must be to transfer knowledge about software and service development from developers to a new generation of young researchers so that they can be active users and not just consumers of infrastructure technology.

Focus: The focus is (1) on informing the participants about the essential pillars of an infrastructure, their structure and the ways to access them; (2) on discussing essentials about web services and the type of interface technologies such as WSDL and REST; (3) on methods to register web services; (4) on providing a programming interface to access a resource; (3) on developing a small application that makes use of web-accessible resources and integrating it into the infrastructure. Thus this summerschool is focusing on showing young people how to actively contribute to the emerging infrastructure and how applications can make use of existing services.

Participants: We expect young PhDs and/or Postdocs who are designing resources and applications to become part of a cyberinfrastructure scenario. We expect some basic knowledge about programming languages such as C# and/or Java and knowledge about XML technology. Thus we address those who see themselves as active contributors to the LRT cyberinfrastructure domain.

Procedure: The participants will get a certain task to be solved that will require to integrate a few resources into the registered domain and to write a small application that also needs to be integrated into the registered domain.

Teachers: Some web services experts of MPI will give the courses together with a set of specialists who have deep knowledge about all relevant aspects of web services standards, registries and technologies.

ANNOTSCHOOL: CLARA Summer School in Semantic and Nonverbal Corpus Annotation and Evaluation (UCPH and MPI)

Goal: Young researchers will be trained within different aspects of semantic and nonverbal corpus annotation as well as in evaluation methods of these. They will get an overview of applicable annotation methods and tools and hands-on skills on a selected set of tools.

Focus: Semantic annotation schemes and annotation of non-verbal behavior, as well as evaluation methods for annotation systems.

Participants: Approx 5-7 researchers within the CLARA network and some external participants.

Procedure: A mixture of theoretical lectures and hands-on exercises. All participants will be trained in both semantic and nonverbal annotation schemes, but one or two course day during the week the participants will get the opportunity to focus on their preferred area of annotation. Tools: GATE, ELAN, ANVIL, and others.

Teachers: Teaching will be held by UCPH staff as well as by two invited international experts in the field. Suggested experts (to be confirmed): Paola Monachesi (Utrecht U), Martha Palmer (U Colorado, Michael Kipp (DFKI), Jens Alwood (Göteborg), and Brian MacWhinney (Carnegie Mellon U).

NEWDEVSCHOOL: CLARA Winter School on New Developments in Computational Linguistics (CUNI)

Goal: ESRs as well as Computational Linguistic masters students will be exposed to recent advances in Computational Linguistics.

Proposed topics and speakers: (a) Statistical parsing for language understanding by Prof. Charniak, head of BLLIP, laboratory for language processing at Computer Science, Brown University, Providence, RI, USA. He is interested in parsing (shallow and deep) for quite some time. He will give present recent hypotheses about contributions to parsing accuracy (b) Lexical resources for language understanding by Prof. Palmer, University of Colorado at Boulder, who is well-known for the creation of PropBank, a lexical resource linked to the world-famous Penn Treebank for verbal argument markup and sense disambiguation. She will present recent results in creating VerbNet and merging various lexical resources (PropBank, FrameNet and others) to a unified, high-quality lexical resource. Important issues regarding the sense granularity of such a resource will also be dicussed.

Time and format: The winter school is organized as a week-long event.

CAREERSCHOOL: Industrial Career Training Course: Product Planning for Next Generation Information Access Technology Solutions (UIB, COMPERIO and FFZG)

Goal: This training course is aimed at providing the next generation of scientists with the complementary skills that are necessary to move from theory to a marketable products and solutions. The course will include planning, market analysis, entrepreneurship, exploitation of research results, project management, proposal writing, communication, research ethics and IPR management.

Focus: The course will both demonstrate a complex industrial system based on multiple components and give an introduction to strategic product planning covering the span from idea to deployment. FFZG will present the case study of the system for semantic analysis of newswire texts that is being developed for Croatian News Agency (HINA). This case study demonstrates how a complex system can be build from simple existing modules such as lemmatization, POS/MSD tagging, named entity recognition and classification, document classification, keyword extraction; with perspective to widen the customer support with more advanced Knowledge Technology such as social network analysis, event detection, trend detection etc. The second part of the course starts with an overview of the principles of market and competitor analysis, followed by a session focused on project management. Basic concepts and benefits of agile development strategy (Scrum methodology) will be demonstrated by examples from industrial projects. The strategy of customer-driven innovation will be explained in detail, and methods will be taught to package experience and Best Practices in product development. The workshop will also discuss a selection of architectural issues, such as best practices in the implementation of clean, backward-compatible and standard-compliant programming interfaces (APIs); as well as the design of service-oriented architectures tailored to the next generation of information access technology solutions. Further modules will include complementary skills such as proposal writing, management, communication skills and ethics.

Organization, time and location: A four day course in Dubrovnik, Sep. 2011, organized by FFZG. The industrial associated partner COMPERIO will be the main responsible for the programme of the course; both partners will draw extensively on their industrial experience to make this workshop a cutting-edge event.

B.1.4 Management structure and procedures

Management structure

The distribution of the work follows from the project descriptions and training actions where for each project and each training action it is indicated who has the responsibility. In the case of shared responsibilities, it is the first partner listed who is the chief responsible for the task.

All partners enter into a consortium with a management structure that is simplified from the DESCA model consortium agreement to fit an Initial Training Network. The organisational structure of CLARA shall comprise the following management bodies and designated persons:

- *General Assembly* is the ultimate decision-making body of the consortium and shall consist of one representative from each level 1 and 2 participant, and one representative chosen by vote amoung all active ESR/ERs. The General Assembly can adjust the project plan by a majority of votes. The General Assembly will have at least a startup meeting and a midterm meeting.
- A *Work Package Committee* will be established for each Work package and will be chaired by the Work package leader; other members will consist of one representative for each other level 1 and 2 participant with planned activities in the Work package, and one representative of the ESR/ERs chosen by vote among the ESR/ERs appointed in the Work package. The Work Package Committees will have at least a startup meeting and a midterm meeting and will in addition keep regular contact by electronic means.
- *Network Coordinator* is in charge of scientific coordination and finances, and is chair of the General Assembly. The network coordinator reports to the General Assembly and supplies all necessary information to the Supervisory Board that the board needs to fulfill its task.
- *Project Manager* is appointed to assist the Network Coordinator in administrative, financial and practical planning tasks, and will provide management support for the organization of meetings, the writing of reports, the midterm review and dissemination activities.
- *Training Coordinator* is in charge of the overall coordination and harmonization of CLARA training activities.
- *Supervisory Board* (see members below) will monitor and advise the consortium on actions related to scientific directions, user needs, professional training, gender issues, community policy and coordination with international research activities. The Supervisory Board will deliver two short reports, one prior to the mid-term evaluation and one prior to the final report. The Supervisory Board will be present at the General Assembly's mid-term meeting.

Internal communication

Apart from the meetings planned for the management bodies, communication between those bodies will otherwise be supported by phone, email, video-conferencing and web-based communication and information. Communication with the ESR/ERs will partly be through their representatives in the General Assembly and Work Package Committees, but also through web-based bulletin boards and newsletters and through an email list. This communication will not be one-way but two-way and the fellows will be involved in setting up and maintaining the bulletin board and newsletter.

Financial management

The overall financial administration will be hosted at the coordinator. A Consortium Agreement will regulate the financial management among the following main lines. Funds for fellows' salaries will be distributed to each partner from the start of each fellow's project and appointment will be done at the partner hosting the project. Funds for project costs that were originally allocated to contractors, not used in due time within the contract period, may be used for readjusting budgets of other contractors that have justifiably higher than anticipated project expenses. Partners agree to contribute to network-wide costs in proportion to their allocated funds, in particular from categories E and G, to co-fund the consortium meetings and joint training programme and to allow the appointment of a Project Manager and Training Coordinator by the Network Coordinator.

Coordination of the training program, announcement of vacancies and appointment of ESR/ERs

The overall coordination of CLARA-wide training activities is the task of the *training coordinator*, who will establish procedures for securing local *supervision* and training as well as *participation* of ESR/ERs in the joint training program. The training coordinator monitors the ESR/ERs training conditions and progress and reports on to the general assembly, which in turn may advise the network coordinator on policies. The actual *organization* of workshops and local training activities is the duty of the respective network partners that host the activities (see esp. B2).

Each partner with ESR/ER person months is responsible for drafting a job announcement for these positions, taking into account eligibility rules and other relevant project information. The work package leader and the network coordinator must approve the announcement before it is placed on the usual channels (Cordis and email lists). Job applications are reviewed by at least two persons designated by the work package leader and from at least two different partners in the network. The network coordinator will perform a formal check on eligibility and general qualifications before the researcher is appointed.

Recruitment of researchers, their appointement and their working conditions will adhere to *The European Charter for Researchers* which specifies the roles, responsibilities and entitlements of researchers as well as of employers and/or funders of researchers, and *The Code of Conduct for the recruitment of researchers.*

Supervisory Board

The members of the supervisory board are selected for their research experience in the subject area, their experience in project management, their experience in training and career development and their links to industry. Should a member of the advisory board resign from his/her position he/she will be replaced by a person of equivalent experience or position. In this case the general assembly suggests candidates for approval by the chair of the advisory board. The following persons have agreed to be on the supervisory board:

- Dr. Laurent Romary, INRIA, is scientific adviser to the Research Department for linguistic computer science and scientific and technical information. He is also director of the newly formed Max Planck Digital Library.
- Prof. Dr. Josef van Genabith is the director of the Centre for Next Generation Localisation (CNGL), an Industry-Academia partnership funded by the Science Foundation Ireland and Industry Partners (2007-2012).
- Dr. Adam Kilgarriff is director of Lexical Computing Ltd, a company offering advanced lexicographic tools, consultancy and corpus-development services.
- Dr. Khalid Choukri is CEO and managing director of ELDA. Incorporated as a company, ELDA is ELRA’s operational body.
- Daniel Ridings is researcher at Mikro Værkstedet, an SMB with a strategic focus on educational materials based on language technology.

B.2 Implementation

B.2.1 Planning of work packages, milestones and deliverables

List and schedule of milestones

Milestone no.	Milestone name	WPs no's.	Lead beneficiary	Delivery date from Annex I	*Comments*
1	TERMCOURSE	8	NHH	9	Jun 2010, Bergen, 35 ERD
2	WORDNETCOURSE	8	UIB	9	Jun 2010, Bergen, 15 ERD
3	LRSCHOOL	8	MPI	10	July 2010, Nijmegen, 25 ERD
4	MORPHCOURSE	8	UHEL	12	Sep 2010, Budapest, 10 ERD
5	EVALCOURSE	8	UCPH	12	Sep. 2010, Paris, 20 ERD
7	TREECOURSE	8	UTU	15	Dec. 2010, Prague, 16 ERD
8	INFRASCHOOL	8	MPI	22	July 2011, Nijmegen, 25 ERD
9	ANNOTSCHOOL	8	UCPH	23	Aug. 2011, Copenhagen, 30 ERD
10	CAREERSCHOOL	8	UIB	24	Sep. 2011, Dubrovnik, 12 ERD
11	NEWDEVSCHOOL	8	CUNI	27	Dec. 2011, Prague, 20 ERD

Tentative schedule of project reviews

Review no.	Tentative timing, i.e. after month X = end of a reporting period	*planned venue of review*	*Comments , if any*
1	Mid-term review after month 24	Bergen	

Work package list

Work package No	Work package title	Type of activity	Lead beneficiary No	Person-months	Start month	End month
1	Consortium Management, Linking, Dissemination and Exploitation	MGT	UIB	0	1	48
2	Designing and Testing Common Infrastructures	RTD	MPI	72	1	48
3	Lexical Semantic Modeling	RTD	UiB	108	1	48
4	Next generation Domain Modeling	RTD	NHH	60	1	48

5	Multimedia and Multimodal Communication Modeling	RTD	UCPH	60	1	48
6	Applications	RTD	UTU	102	1	48
7	Parsing Technologies and Grammar Models	RTD	CUNI	168	1	48
8	Joint Training Programme	RTD	UIB	0	1	48
	TOTAL			570		

Notes: Person-months in the above tabel are ESR and ER only. In addition, WP3 has 2 VS months, WP4 has 3 VS months and WP7 has 3 VS months. Efforts related to management and teaching are not included.

List of Deliverables – to be submitted for review to EC

Del. no.	**Deliverable name**	**WP**	**Lead beneficiary**	***Estimated indicative person-months***	**Nature**	**Dissemination level**	**Delivery date (month)**
1	REPORT1	1	UIB	0	R	PU	12
3	REPORT2	1	UIB	0	R	PU	24
3	REPORT3	1	UIB	0	R	PU	36
4	JOINT TRAINING PROGRAMME	8	UIB	0	O	PU	36
5	REPORT4	1	UIB	0	R	PU	48
6	PROJECT2A	2	MPI	36 (MPI)	O	PU	48
7	PROJECT2B	2	MPI	36	O	PU	48
9	PROJECT3B	3	UIB	36	O	PU	48
10	PROJECT3C	3	UCPH	36	O	PU	48
11	PROJECT3D	3	UPF	36	O	PU	48
12	PROJECT4A	4	NHH	30 (NHH) + 30 (TILDE)	O	PU	48
13	PROJECT5A	5	MPI	24	O	PU	48
14	PROJECT5B	5	UCPH	36	O	PU	48
15	PROJECT6A	6	UTU	24 (UTU)	O	PU	48

16	PROJECT6B	6	TILDE	21	O	PU	48
17	PROJECT6C	6	TILDE	21 (TILDE) + 36 (CUNI)	O	PU	48
18	PROJECT7A	7	UHEL	48 (UHEL) + 24 (UTU)	O	PU	48
19	PROJECT7B	7	UIB	36	O	PU	48
20	PROJECT7C	7	CUNI	36 (CUNI) + 24 (UTU)	O	PU	48
TOTAL				570			

Notes: Person months in the above table are ESR and ER only. In addition, PROJECT3C has 2 VS months, PROJECT4A has 3 VS months and PROJECT7C has 3 VS months.

Table of work package descriptions

Note: all Work packages start in month 1 and end in month 48, unless indicated otherwise.

Work package 1: Consortium Management, Linking, Dissemination and Exploitation

Lead Participant: UIB.
Other participants: NHH.
Person-months: No ESR/ERs are allocated to this task.

Objectives:

This work package is aimed at the implementation of the management structure, further explained in section B1.4, at linking to users and relevant research groups, and at coordinated dissemination and exploitation of the results.

Description of work:

Work package activities consist of overall management of the project and the training activities, writing of reports, consortium meetings and the planning of dissemination and exploitation. Dissemination activities include presence at conferences, web-based targeting of relevant user groups and contacts with other relevant projects, in particular with the CLARIN project.

Deliverables:

REPORT1, REPORT2, REPORT3 and REPORT4 (see List of Deliverables) are the collections of reports due after years 1, 2, 3 and 4 respectively, including periodic reports, midterm review and final report. These reports will include progress on the scientific work, on training and on dissemination and exploitation. For a further description of these reports, please refer to the FP7 Guidance Notes on Project Reporting.

Work package 2: Designing and Testing Common Infrastructures

Lead Participant: MPI.
Person-months: MPI 72 ESR.

Objectives:

This work package is aimed at the design of the next generation of common language resource infrastructures on which future models will be based. It will have a strong link with CLARIN to reach its goals. The following research topics can be discerned:

- sustainability and long term preservation of language resources
- construction of an interoperable language resource infrastructure
- efficient management of and access to very large language resource repositories
- organizational and technical models for large service grids
- design of a discipline-oriented service architecture
- development of new standards and promotion of existing ones

Deliverables with description of work:

Project 2A: Common Language Resources Infrastructure (MPI)

Objective: We propose to build exemplaric modern online repository with structured metadata descriptions and relevant linguistic content. Web applications and services will be used and extended to allow the manipulation and usage of resources such as annotations and lexicons via the web.

Research strategy and methodology: Relevant resources from different creators that used different linguistic encoding standards are analysed, described with metadata, transferred into archivable formats and then integrated into such an archive system. This will include the resources of the TRACTOR archive of Central and East European language resources, collected as part of a major European initiative, but currently unavailable to researchers. Applications will be written to harvest metadata from different centers to create an integrated browsable and searchable domain. Web applications will be created to create virtual collections by making use of federation middleware so that simple operations can be carried out on the virtual collection.

Feasibility, innovative aspects and relevance: The results can immediately be integrated into the CLARIN infrastructure, the technologies will conform to emerging GRID standards, and the project will make use of existing tools resources already held at MPI. The project is fully scalable in terms of the scope of resources to be included.

ESR training aspects: Training will be provided in principles of digital archiving and metadata standards, and in relevant tools.

Project 2B: Interoperability Standards and Semantic Web Technology (MPI)

Objective: We propose to tackle interoperability problems at the syntactic and semantic level. Annotation structures and lexicons will be analysed and transferred to generic schema-based formats such as LMF, XCES and EAF. The resulting resources will be integrated into repositories such that joint operations are made possible. The expected semantic interoperability problem will be tackled by integrating all linguistic concepts into a concept registry based on the ISOcat model and appropriate relations will be registered allowing cross-resource searching for example.

Research Strategy and methodology: Resources in legacy formats will be converted into generic schemas, taking into account structural and character encoding details, probable information losses and semantic interpretation aspects. The transformed resources are integrated into a repository infrastructure that allows virtual collection building. For the created virtual collections an analysis of the linguistic concepts is made. For those that match the official ISOcat definitions, references are created in a practical ontology, for those that are not contained entries are made in a project registry instance. An small RDF framework is set up to create relations between a selection of concepts so that joint search operations can be carried out exploiting the registered relations and concepts. A web application is created that allows to store and therefore re-use the elements of the created practical ontology.

Feasibility, innovative aspects and relevance: The generic formats have been tested against a wide number of structures. The ISO DCR model is stable, the ISOcat software framework has been demonstrated at ISO meetings and the API is already in use by web-applications so that there is a running concept registry framework. For creating relations tools have been developed by the semantic community which need to be extended to fit to the needs of the project. A search engine can be developed reading the resources being represented in generic formats and exploiting the ontology specifications.

Work package 3: Lexical Semantic Modeling

Lead Participant: UIB.
Person-months: UIB 36 ESR, UCPH 36 ESR, UPF 36 ESR.

Objectives:

This work package is aimed at state-of-the-art semantic modelling, consisting on the one hand of the exploration of bilingual corpora (LC and UIB) towards computer-assisted lexicography; on the other hand, word nets will in turn be used to enhance the information in corpora through advanced annotation schemes (UCPH). There will be secondments between UPF and TEMIS.

Deliverables with description of work:

Project 3B: Exploring bilingual corpora with Semantic Mirrors (UIB).

Objective: Lexical semantic databases and word nets have been manually constructed for a number of languages, but are still missing for several smaller languages. We propose to construct experimental word net-type lattices and thesauri that distinguish and interrelate word senses of homonymous and polysemous words) for European languages which have not yet been researched with the Semantic Mirrors method, and thus to contribute to the development of semantic applications for the languages in question.

Research strategy and methodology: Translation corpora are first lemmatized and aligned at sentence level (using e.g. TCA2) and word level. The set of possible translations of each lemma in both languages is collected, constituting the 'translational image' of the lemma. From overlap relations and correspondences among the translational images, word senses are individuated, and 'semantic fields' revealing assumed hyperonymy/hyponymy and near-synonymy relations among the senses are computed and represented as complex semilattices. Thesaurus-type databases, including further subsense individuation subject to certain set parameters, are finally derived from the semilattices. The results will be compared to results of other method of sense individuation, in particular the Sketch Engine based method (LC).

Feasibility, innovative aspects and relevance: The identification of word senses is highly relevant for applications in the area of web search, information extraction and dictionary production. The possibility of partly automatizing the development of such resources is therefore extremely desirable. Resources derived from translational corpora by means of the Semantic Mirrors method will be especially relevant for multilingual language processing.

ESR training aspects: Training will be provided in parallel corpora, alignment tools and the Semantic Mirrors method.

Project 3C: Semantic annotation of large corpora (UCPH)

Objective: Semantically annotated corpora, also for the lesser studied languages, may well prove to be a necessary prerequisite for moving into the next era of language technology where more sophisticated semantic web-based tools for i.e. semantic analysis and data mining can be developed. We will investigate (i) to which degree current wordnets (currently available for around 60 languages) are suitable as lexical semantic resources for automatic annotation of large, varied corpora (written, spoken and visual) and (ii) to which degree resources and tools for corpus annotation are transferable between languages.

Research strategy and methodology: As a necessary prerequisite for supervised annotation systems, the suggested methodology includes manual semantic tagging of training corpora for the relevant languages to be studied, including, however, investigations of whether the building of such Gold Standards can be supported by means of parallel corpora as envisaged in the Semantic Mirror or Sketch Engine methodologies (see other projects in this Work package). The Gold Standards will function as training corpora for machine learning systems. Based on Senseval results, which have now been achieved for series of different languages, as well as on the results of Semantic Mirrors methodology which has a cross-lingual, corpus-based focus, the research project will include the study of cross- and multilingual aspects of semantic corpus annotation and the possibility of technological transfer between languages.

Feasibility, innovative aspects and relevance: Semantic corpus annotation is moving into a new era where also lesser studied languages can provide the necessary background language resources required for the task. Semantic corpus annotation can provide the much wanted synergy between large text repositories and lexical-semantic resources and thereby make text collections more accessible for different kinds of humanities researchers.

ESR profile and training aspects: Training will be provided in semantic annotation and word net technology.

Project 3D: Automatic lexical acquisition (UPF)

Objectives: A large number of language models and applications (e.g. machine translation, question answering, etc.) are facing a bottleneck, as they need large lexical resources which are costly to build and maintain for every language and for every new domain. The goal of our research is to provide the automatic acquisition of lexical information to be included in computational lexica, terminology, thesauri, etc. Systems that reduce the cost of producing lexical resources will contribute to reduce the distances among language communities. Our research will work on methods and techniques that ensure portability in multilingual models and applications.

Research strategy and methodology: The performance of different techniques (including Bayesian classifiers, Decision Trees, Hidden Markov Models, Support Vector Machines) will be compared to study the class of learners that better fit the characteristics of linguistic data. The work will adhere to proposed standards for lexical information encoding such as LIRICS Lexical Markup Framework (now under ISO/TC 37/SC 4) and Data Category Registry while paying prioritary attention to lexical resources used by CLARA industrial partners.

Feasibility, innovative aspects and relevance: Current approaches do not immediately handle two crucial problems we should critically address: lexical ambiguity and missing information, which are intrinsic characteristics of natural language. These will be addressed in this project.

ESR profile and training aspects: A candidate with a background including statistics and machine learning will be introduced to lexicology, or vice versa.

Secondment to 3D: Large-scale Information Extraction (TEMIS)

In order to benefit from methods and tools available at the industrial partner TEMIS, the ESR in 3D will spend a short period (dates to be decided later) at TEMIS and obtain training related to the following subproject:

Objective: The application of information extraction approaches on real-world tasks emphasize a number of parameters, that are often different from more research-oriented contexts. Ease of deployment, resource efficiency or the level of maturity for instance can all turn out as descives factors and can influence descisions more than small differences in accuracy. A focus on the assessment of these factors and their consequenences for practical information extraction is of high relevance for an industrial approach. Parameters that enter such considerations may be the availability of a training corpus, the specificity of the information which needs to be processed, the desired level of accuracy, time schedule, and many others.

Research startegy and methodology: In this programme selected information extraction tasks will be considered and investigations will explore various alternatives for their implementation that take into account the above-mentioend parameters. We will aim at making steps towards a methodology that allows to classify a task together with its side conditions (see above) according to which appears to be the best approach.

Feasibility, innovative aspects and relevance: The activities will be based on an existing Information Extraction System called Luxid, which gives access to a range of NLP methods such as Part-of-Speech-Tagging, Shallow Parsing or Sequence Labelling with Conditional Random Fields and offers interfaces for the integration of additional ones. Since the system is employed in production environments at large organisations it represents a reasonable testbed for the investigation of parameters that influence the design choices for the implementation of a given information extraction task.

ESR profile and training aspects: Reseachers will be involved in real-world information extraction tasks and the process of deciding on the proper way to address them.

Work package 4: Next Generation Domain Modeling

Lead Participant: NHH
Person-months: NHH 30 ESR, TILDE 30 ESR.

Objectives:

This work package is aimed at the harmonization of terminological resources across Europe.

Deliverables with description of work:

Project 4A: Harmonisation of Terminological Resources (NHH, TILDE, UIB, AKSIS)

Objective: For professional and technical language resources, the harmonisation of language resources across Europe poses problems at both conceptual and technological levels. Although there are a large number of terminological resources in Europe, those resources are fragmented, located in different institutions and in different formats. We propose to develop language-independent methods and tools for constructing and consolidating multilingual termbases and ontologies, for corpus-based term extraction and for ontology-based domain recognition of text. Moreover, we propose to investigate the degree of compatibility of different term bases at the top level of domain classification, at the mid-level of sub-domain, and at the lowest conceptual level, i.e. the concept itself and its hierarchical relations to neighbouring concepts within the sub-domain. The project aims to develop new, innovative approaches to cross-termbase integration.

Research strategy and methodology: The training will be organized at NHH (with the cooperating partners UIB and AKSIS) and TILDE in order to offer ESRs training on different aspects of terminology work and paying special attention on semantic enrichment and integration of terminological resources. Domain-specific resources are used iteratively to identify sections of the general corpus with a high content of text from the given domain, to provide an initial classification of the general corpus. New domain specific term candidates are extracted semi-automatically from the machine-classified texts, which are in turn fed into the pool of domain-specific resources after manual check. Finally, term candidates undergo a concept-based match with terminological resources representing other languages and are integrated into a common multilingual termbase which is compiled in accordance with accepted metadata standards (IMDI).

Feasibility, innovative aspects and relevance: The proposed methods, tools and domain-specific terminologies are beneficial to end-user systems like machine translation, text summarisation, text generation etc. Automatic domain recognition of text may also improve the precision of taggers and parsers, word sense disambiguation, etc. The online terminology data bank facilitates terminology data accessibility and exchange, standards of EuroTermBank project facilitates terminology interoperability, sharing and reuse. Finally, multilingual term bases are crucial for efficient (manual) translation and instrumental to the preservation of European cultural-linguistic heritage and as a preventive against domain-loss in smaller languages.

ESR training aspects: Training will be provided in term base technology, term extraction methods, term classification approaches, ontology construction, metadata and annotation schemes, industrial best practice, and will include a secondment.

Secondment to 4A: Investigation of (semi) automatic import strategies for existing thesauri and terminologies (TEMIS)

In order to benefit from additional training in methods and tools at the industrial partner TEMIS, ESRs will be detached at TEMIS for the following subproject (dates to be decided later).

Objective: A critical issue for the efficient deployment of methods of information extraction/automatic indexing (IE) is the rapid definition of analysis grammars and rule sets. Ideally existing thesauri can be used, but these are often not directly useable for automatic processing. Using appropriate heuristic and linguistic methods that take into account the structure of both the existing terminologies as well as the target representation of the NLP grammar is decisive to make this process manageable.

Research Strategy and Methodology: In this subproject the task to define and apply appropriate linguistic and heuristic methods to existing thesauri and terminologies will be studied using a finite-state information extraction platform as a reference.

Feasability and innovative aspects and relevance: The activities will be based on an existing Information Extraction System called Luxid, designed by TEMIS.

ESR training aspects: Reseachers will be involved in real-world information extraction tasks and the process of deciding on the proper way to address them.

Work package 5: Multimedia and Multimodal Communication Modeling

Lead Participant: UCPH
Person-months: UCPH 36 ESR, MPI 24 ER

Objectives:

The study of human verbal behavior has long been isolated from nonverbal ways of expression even if they are inherently linked. In order to study human communicative behavior empirically, we need to build multimedia and multimodal resources, requiring new annotation schemes to be designed and applied. The projects below complement each other. The UCPH multimodal corpus and the connected annotation scheme for non-verbal behaviour can be used as testbed for the MPI research, while the results at MPI will verify if the UCPH categories are useful to classify regular patterns.

Deliverables with description of work:

Project 5A: Multimedia and Multimodality Resources and Technology (MPI)

Objective: We propose to build smart pattern recognition components and integrate them into a widely-used multimedia annotation framework so that new layers of annotations on sound and/or video recordings are created which can then be exploited by existing flexible search frameworks. The modules can be of overseeable complexity such as a silence/signal detector or more complex by cascading.

Research strategy and methodology: Typical audio and video recordings are studied to detect regularities and to then allow the participants making a formal description of the patterns over time. These formal descriptions are transferred into a software component that can be integrated into the existing ELAN framework and that creates an additional layer of annotation. It will be investigated for each component whether the existing flexible search engine needs to be extended to allow the combination of the created annotations with the existing ones to yield the expected recognition results. In particular we foresee the extension towards stochastic and learning methods.

Feasibility, innovative aspects and relevance: The paradigm has been investigated already by senior researchers and essential steps for a flexible extendible framework have been made. Different modules can be combined and their effects and combination with other patterns will lead to new insights and will inspire theorizing which will be an important step to overcome the limitations of purely stochastic systems such as Hidden Markov Models or Artificial Neural Networks.

Project 5B: Empirical Studies of Multimodal Behaviour (UCPH)

Objective: One line of research in multimodal behaviour individuates non-verbal patterns such as the detection of eye, hand and body movements and trains algorithms to interpret these movements. Another line of research aims at recognizing the communicative functions of non verbal behaviours and studying the interaction of these behaviours by developing multimodal corpora, creating annotation schemes and building up reliably annotated multimodal corpora covering different cultural and social settings. The present project has the goal to produce insight on essential aspects of multimodal behaviour such as feedback, dialogue management, information structure and reference by conducting empirical studies in audiovisual corpora based on manual annotation and analysis as well as automatic methods such as machine learning.

Methodology: Existing multimodal corpora as well as multimodal data annotated within the project will be used to produce empirically-based evidence of the way in which specific non-verbal behaviors are used in human communication. Relevant research issues are for example to which extent this behaviour may be captured in formal categories that an algorithm can be trained to recognize, how crucial the cultural context is to interpret the behavior, or whether a universal core can be defined for non-verbal behaviours, similarly to what certain linguistic schools have attempted to do for language. Analyses of multimodal communication in different cultural settings, communication situations and social contexts are also relevant.

Work package 6: Applications

Lead Participant: UTU.
Person-months: UTU 24 ESR, TILDE 42 ESR, CUNI 36 ESR.

Objectives:

This work package is concerned with ICT applications of next generation language models in key sectors: the educational sector and the language services sector (translation, proofing, information retrieval).

Deliverables with description of work:

Project 6A: Connecting ICALL Research and Development to Foreign Language Teaching Needs (UTU)

Objective: The project will investigate and bridge the gap between research in Intelligent Computer-Assisted Language Learning (ICALL) and the actual needs and practice of current Foreign Language Teaching (FLT) – a crucial step towards the next generation of intelligent FLT tools.

Research strategy and methodology: The gap between ICALL research and FLT practice is explored by (a) identifying needs and shortcomings in FLT practice, (b) reviewing the results from Second Language Acquisition research, in particular concerning the role of noticing and awareness of language forms and of feedback, (c) designing activity types which can fill those needs, (d) designing the language model, activity model, and learner model, (e) developing and reusing NLP approaches and resources, such as those developed in the network, to analyze and provide the individualized feedback to the learner, and (f) evaluating the resulting ICALL activities with real-life learners. The analysis of learner input in ICALL systems is also closely related to the annotation of learner corpora (cf., e.g., the recent CALICO workshop on the topic), for which the annotation tools and resources developed and made available in the network are directly relevant.

Feasibility, innovative aspects and relevance: The research and training components of this project can rely on a concrete starting point, the TAGARELA system, which Detmar Meurers and Luiz Amaral developed and successfully used in Portuguese courses at the Ohio State University. On this basis, the project can successfully investigate which activity types needed in FLT practice can be integrated, where the necessary NLP analysis is available and effective, and how explicit learner modeling can improve the disambiguation of the learner input.

Project 6B: Proofing tools and resources for under-resourced languages (TILDE).

Objective: We propose knowledge transfer to develop proofing tools (mainly grammar checkers) for languages that currently do not have these technologies available.

Research strategy and methodology: Grammar checkers for world's major languages are developed and included in text processing software. However majority of small languages luck this feature. Grammar checkers are not much researched during the last decade; they typically use shallow syntactic structure rules to find agreement and word order errors and stylistic problems. This approach does not give good results for highly inflected languages with rather free word order. The first research direction will focus on multilinguality of grammar checking techniques. Existing grammar checking techniques will be evaluated to find language specific phenomena and new tools will be developed. The second research direction will be an investigation of the existing grammar checking methods and development of new methods to create suitable techniques to process deep syntactic structures of highly inflected languages.

Feasibility, innovative aspects and relevance: Good proofing tools are important for any language to survive and develop, but they are crucial for small and under resourced languages. Grammar checkers for highly inflected languages are not well researched and developed. Europe is multilingual with many languages; research groups working on proofing tools for under resourced languages are quite isolated and scattered through Europe therefore a network for gathered research and knowledge exchange is necessary.

ESR training aspects: Training will be provided in parsing technologies, text processing, CFG and dependency grammars.

Project 6C: Translation tools and resources for under-resourced languages (TILDE, FFZG and CUNI).

Objective: Despite significant progress in machine translation (MT) technologies globally, the circle of languages and language pairs that take advantage of these technologies remains limited. These technologies are not available for a number of languages in Europe, for others they are often restricted to only one language pair – the national language and English. We propose to investigate possibilities to facilitate development of translation tools and resources for languages that currently do not have or have limited translation technologies and resources.

Research strategy and methodology: The first research direction will be an investigation of the existing MT methods to find the most suitable solution (rule-based, data-driven, hybrid) for highly inflected languages with limited or no parallel corpora available. The second direction will be research of innovative content creation methods (mainly parallel corpora, and lexicons) to find effective methods to create content necessary for data-driven MT technologies. The third direction will be the base language technologies and resources (lemmatizer, taggers, parsers and dictionaries) for under resourced and highly inflected languages. CUNI will provide large-scale resources for Czech, an inflective language targeted in this project, and will use a statistical paradigm based on phrase- and tree-based transfer models with linguistic features. This project will be ***lab-based*** as CUNI will use time on its High-Performance Computing cluster.

Feasibility, innovative aspects and relevance: The current research in MT field is mostly English centered and assumes availability of large parallel corpora. We propose to do MT research for highly inflected languages with rather free word order and limited parallel corpora available. Research groups working on MT for under resourced languages are quite isolated and scattered through Europe therefore a network for gathered research and knowledge exchange is necessary.

ESR training aspects: Training will be provided in different translation technologies, parallel corpora, text processing, lexicon building and alignment tools.

Work package 7: Parsing Technologies and Grammar Models

Lead Participant: CUNI.
Person-months: CUNI 36, UHEL 48, UTU 48, UIB 36.

Objectives:

This work package is aimed at the exploration of next generation parsers. New finite state approaches will be developed for the construction of hybrid models. Furthermore, the next generation of analyzers must perform a deeper analysis in order to support meaning extraction. This requires a careful tuning of analyzers to empirical data in treebanks, i.e. syntactically annotated corpora, and research into deep structural commonalities and differences between languages.

Deliverables with description of work:

Project 7A: Finite-state parsing methods (UHEL and UTU)

Objective: Consistent processing the some 100 languages, including European official and otherwise relevant languages is necessary for achieving common European platforms. This requires adequate computational metods and software, solid linguistic principles and common standards. Methods using finite-state transducers (FSTs) have proven to be a general framework for describing morphological and phrase level processing of any European language and provide a technically uniform platform for the implementations. New possibilities for relating the FST algorithms to language models and formalisms for parsers have emerged due to recent innovations in the methods for compilation. The aim of the project is to deepen the understanding of these methods and invent new applications for large scale language processing tasks such as shallow parsing and information extraction.

Research strategy and methodology: One of the aims is to combine rule-based models and probabilistic methods through the use of weighted FSTs. Formal elegance and practical efficiency can be pursued by the composition of various language models once they are represented as weighted or unweighted FSTs. New methods allow for the compilation of a wide variety of constraints. The Generalized Restriction (GR) operator enables the implementation of sophisticated rule-based constraints, and the Optimality Operator allows for the implementation of discrete-valued probabilistic constraints. Both of these rule-based approaches can be combined with traditional statistically based methods, which provide the weights for other parts of models. This mixture of rule-based and probabilistic approaches will provide a powerful and elegant approach to practical problems in NLP.

Feasibility, innovative aspects and relevance: The methods to be created are new, better adjusted for modelling of different aspects of language than those of Xerox XFST or Helmut Schmid's SFST. Still, FST techology is well understood and feasible. The innovations enable an almost infinite range of language models to be expressed in a technically identical form as weighted FSTs which can be combined (and composed) in unrestricted ways. The methods and the resulting FSTs are language independent. Results are readily applicable for SMEs for building commercial modules and researchers for building open source modules.

Project 7B: Next Generation Deep Grammar Models (UIB and AKSIS)

Objective: Commonalities and differences between structural properties in languages have been studied, but not yet by means of a large scale multilingual corpus analysis. We want to determine to what extent the development of parallel deep grammars for typologically diverse languages may support the automatic derivation of high-quality parallel treebanks for the languages, suitable as a basis for a deeper theoretical understanding of the ways in which syntactic functions, semantic roles and translation are interrelated.

Research strategy and methodology: We propose to construct aligned multilingual treebanks as parsed corpora for a number of languages. The corpora are batch parsed by means of the XLE (Xerox Linguistic Environment) and disambiguated by means of the LFG Parsebanker. This project will be ***lab-based*** as it will use time on a High-Performance Computing cluster at Unifob.

Feasibility, innovative aspects and relevance: Parallel treebanks aligned at phrase level are innovative resources for gaining new insights in translational correspondences at structural levels. The project depends on the availability of LFG grammars. A large LFG grammar has been developed for Norwegian in the NorGram project led by Prof. Helge Dyvik, while other LFG grammars are being developed for several other languages in the ParGram project.

ESR training aspects: Training will be provided in large scale grammar development, treebanking, and use of the XLE parser and LFG Parsebanker tools.

Project 7C: Linguistic Analysis for Treebank Annotation (CUNI and UTU)

Objective: We propose here to specify a common core of syntactic and semantic features for a range of languages (Czech, English, German, Slovak, and others). The development of a common core of such categories and features will greatly facilitate the comparability of treebanks for different languages and overcome the current state of the art, where the lack of comparability of such resources has been recognized as a serious problem for theoretical and computational research. We will test and validate the resulting annotation schemes by performing annotation of texts for these languages, providing as the final product annotated corpora (treebanks) for both linguistic research and statistical language learning (i.e., creating tools for automatic syntactic and semantic analysis).

Research strategy and methodology: A corpus is morphologically analyzed and pre-parsed by state-of-the-art disambiguation and parsing tools. At CUNI, the intermediate data is then manually disambiguated by the annotators (linguists, mainly) using the TrEd graphical UI. At UTU, the Annotate tool is used to semi-automatically select the correct analysis for a given sentence.

Feasibility, innovative aspects and relevance: A highly relevant research issue concerns the synthesis between dependency-based and constituency-based annotations. Bringing together two prominent institutes in the two respective traditions will provide a good basis for an innovative solution to this important desideratum. The project is imminently feasible given that the manual and automatic tools are already available or can be easily adapted. At CUNI, the PML is easily adaptable for novel linguistic phenomena to be specified and annotated by the ESRs under the supervision of the CL professors at CUNI. Then, new tools can be developed or existing tools improved by machine learning based on the annotated corpora. At both CUNI and UTU, language independent tools for annotation and error detection can easily be applied to new languages and the adapted annotation schemes to be developed.

Work package 8: Joint Training Programme

Lead Participant: UIB
Person-months: 0 (No ESR/ER will be allocated to this work package, but ESR/ER/VSs will participate in the activities).

Objectives:

This work package will provide a joint training program consisting of thematic training courses and summer schools, targeted at ESR/ERs in the network as well as at dissemination to external researchers, especially those at CLARIN members.

Deliverables with description of work:

TERMCOURSE: Thematic Training Course on Methods and Technologies for Consolidating and Harmonising Terminological Resources (NHH, TILDE, UIB and AKSIS)

Goal: The workshop is aimed at providing the ESR/ERs the necessary skills to utilize existing language resources in new, innovative ways, for the overall purpose of harmonizing Europe's terminological resources. The main aim is to ensure the transfer of knowledge from staff within and beyond the CLARA network working on relevant R&D projects, in order to enable participants to specify aspects on which to focus their own research efforts.

Focus: The focus will be on theoretical and technical aspects, and ESRs/ERs will be given advanced training in topics relevant for pan-European integration of terminological resources. Theoretical aspects include a) the organization of knowledge and its representation in structured databases, b) classification of domains and sub-domains, c) methods for investigating domain classification across term bases. Technological aspects include a) knowledge of existing term bases and their formats, b) knowledge of existing terminological standards and metadata schemes; c) corpus-based term extraction technology; d) multilingual integration and interoperability.

Participants: Given the interdisciplinary nature of the workshop, the course will be relevant for PhD scholars and/or postdocs who have a basic training in a relevant theoretical or technical field (terminology, linguistics, knowledge management, computational linguistics, termbase technology, language resources) but who need to increase their knowledge in relation to computational methods in terminology work.

Procedure: The course will be given as a combination of lectures, group tutorials and hands-on training in the use of resources and technology.

MORPHCOURSE: Thematic Training Course on Processing Morphologically Rich Languages (UHEL and HASRIL)

Goal: Morphologically-rich languages like Turkish, Finnish, Hungarian, etc., present significant challenges for natural language processing applications due to their relatively free word order and highly productive morphological processes (inflection, agglutination, compounding). The course will work on problems due to dictionary size, sparse data, poor language model probability estimation, high out-of-vocabulary rate and information gaps on related lexical items.

Focus: The course will introduce advanced modelling techniques addressing these problems, such as decomposition of complex word forms into smaller units, relating inflectional variants to root forms (lemmatization), methods for optimizing the selection of units at different levels of processing, novel probability estimation techniques, and the creation of a new class of data resources and annotation tools. Newer finite state techniques have proved to be useful and can be optimized in combination with other methods. The course will conclude with an assessment of present day standard techniques and a demonstration of practical applications focusing primarily on Hungarian and Finnish.

WORDNETCOURSE: Thematic Training Course: Ontologies and Wordnets and Their Use in NLP Technologies (HASRIL, UIB, LC and UCPH)

Goal: In recent years both wordnets and more formal ontologies have grown to be crucial background databases for various applications. WordNets are being used in word sence disambiguation, machine translation, information extraction and information retrieval, just to list some application areas. Over 60 wordnets have been developed over the world. Languages that are typologically different than the main model language, English, had to face additional linguistic tasks when constructing their semantic network. With theoretical focus on representing verbal structures in a lexical semantic network, the Hungarian WordNet was the first one to deal with lexicalised event structures in a systematic way. For some languages, wordnets do net yet exist and new compuational methods (such as Semantic Mirrors) are being explored to support their creation.

Focus: The course will, make a clear distinction between formal and linguistic ontologies, and give a theoretical overview of general questions concerning WordNet-building, such as a comparison of different methods (fully automated and semi-automated methods). Problems of languages that may express event structure through lexical means, e.g. through prefixes sensitive to aspect and Aktionsart, will be highlighted, and NLP applications relying on WordNets in general as well as the verbal WordNet, will be discussed in order to give a taste of practical issues. Finally, methods for automatically creating and exploring lexical relations, including SketchEngine and Semantic Mirrors, will be taught.

TREECOURSE: Thematic Training Course on Methods and Technologies for Consolidating and Harmonising Treebank Annotation (CUNI and UTU)

Goal: The course focuses on two aspects: the resources and tools as well as the linguistic analysis underlying the annotation schemes of treebanks. It will support the exploration and comparison of dependency-based and constituency-based annotation schemes and will provide the ESRs with the necessary skills to utilize existing resources and tools for treebank annotation and to adapt the them to the revised and harmonized annotation schemes.

Focus: The focus is on the comparability of existing annotation schemes for treebanks across a number of languages. It adresses both the analytic basis and the prospects for harmonizing annotations across different languages and across different linguistic frameworks.

Participants: The event will bring together the ESRs from CUNI and UTU, but will be open to all interested ESRs, especially those dealing with annotation of corpora, and researchers from other CLARA training sites.

Procedure: The course will be given as a combination of lectures, group tutorials and hands-on training in the use of resources and technology.

EVALCOURSE: Thematic Training Course on Evaluation of Human Language Technologies (ELDA and UCPH)

Goal: For any HLT research effort to be successful, it is essential that it be assessed through rigorous evaluations of the developed technologies. This allows performance benchmarking and a better understanding of possible limitations and challenging conditions. The workshop aims at providing ESRs/ERs the background and skills to use and implement state-of-the–art evaluation tools and techniques for speech technologies, grammars and parsing, machine translation and speech-to-speech translation, information retrieval/filtering, multimodal interfaces, etc.).

Focus: The workshop will elaborate on the role of evaluation on the research progress, on the need for a truly European infrastructure for HLT evaluation, on the main reasons to promote an international dimension of the evaluation, insisting on the multilingual issues. The workshop will introduce and describe some evaluation concepts (comparative evaluation versus competition, technology evaluation versus usage/usability evaluation). It will also describe the different types of evaluation and how to ensure that evaluation does not kill innovative not-yet-mature approaches.

Participants: Given the interdisciplinary nature of the workshop, the workshop will be relevant for any ESRs/ERs who are involved in the development of algorithms and systems for speech technologies, machine translation, parsers, information retrieval, multimodal interfaces, etc..

Procedure: The workshop will be given as a combination of presentations, group tutorials and hands-on training in the use of evaluation technologies for HLT.

LRSCHOOL: CLARA Summer School in Advanced Resource Creation, Archiving and Usage (MPI)

Goal: Young researchers will be trained in how to use modern technology to create language resources in particular when the source material are multimedia streams, how the resulting complex resource types can be archived, how they can be accessed via state-of-the-art web applications and how they can be enriched. It will also be shown how virtual collections can be built and how operations can be carried out on such collections. The result must be that young people have a deep understanding about modern methodologies and technologies to create, archive and use sharable resources.

Focus: The focus is (1) on using state-of-the-art tools to create resources that adhere to open standards such as XML and MPEG, (2) on teaching of how to optimally make use of open archives, how to use converters for various data types and how to define the necessary access permissions, (3) on existing frameworks that allow accessing the archived resources via various ways (from metadata up to web applications for complex objects), (4) on existing methods of how to create and use virtual collections and (5) on existing frameworks to make comments and draw relations between resources and resource fragments.

Participants: We expect young PhDs and/or Postdocs who are going to work on/with language resources in their work and who need to be educated to use state-of-the-art tools. We expect some knowledge about computational aspects, but will not require deep skills about XML schema or software programming. Thus we address those who see themselves as being users of modern technology and methodologies.

Procedure: The participants bring own resources with them such that all indicated steps can be carried out using and combining their material. The LAT technology from MPI will be used during the course. Some technology form others will be used to do efficient processing of specific tasks (speech analysis, conversion, etc).

Teachers: Mainly members of MPI will give the courses. In addition we will invite specialists who have deep knowledge about relevant standards in our domain, about audio/video codecs and appropriate software, about speech analysis and appropriate software, about semantic web techniques and representation standards such as RDF. The MPI experts have 8 years of experience in giving such courses twice per year to the indicated group of people.

INFRASCHOOL: CLARA Summer School on Infrastructure Tool Development (MPI)

Goal: Young researchers will be trained in an advanced course about all aspects that are relevant for making use of the emerging CLARIN infrastructure and about how to actively contribute to it with new resources and tools. The result must be to transfer knowledge about software and service development from developers to a new generation of young researchers so that they can be active users and not just consumers of infrastructure technology.

Focus: The focus is (1) on informing the participants about the essential pillars of an infrastructure, their structure and the ways to access them; (2) on discussing essentials about web services and the type of interface technologies such as WSDL and REST; (3) on methods to register web services; (4) on providing a programming interface to access a resource; (3) on developing a small application that makes use of web-accessible resources and integrating it into the infrastructure. Thus this summerschool is focusing on showing young people how to actively contribute to the emerging infrastructure and how applications can make use of existing services.

Participants: We expect young PhDs and/or Postdocs who are designing resources and applications to become part of a cyberinfrastructure scenario. We expect some basic knowledge about programming languages such as C# and/or Java and knowledge about XML technology. Thus we address those who see themselves as active contributors to the LRT cyberinfrastructure domain.

Procedure: The participants will get a certain task to be solved that will require to integrate a few resources into the registered domain and to write a small application that also needs to be integrated into the registered domain.

Teachers: Some web services experts of MPI will give the courses together with a set of specialists who have deep knowledge about all relevant aspects of web services standards, registries and technologies.

ANNOTSCHOOL: CLARA Summer School in Semantic and Nonverbal Corpus Annotation and Evaluation (UCPH and MPI)

Goal: Young researchers will be trained within different aspects of semantic and nonverbal corpus annotation as well as in evaluation methods of these. They will get an overview of applicable annotation methods and tools and hands-on skills on a selected set of tools.

Focus: Semantic annotation schemes and annotation of non-verbal behavior, as well as evaluation methods for annotation systems.

Participants: Approx 5-7 researchers within the CLARA network and some external participants.

Procedure: A mixture of theoretical lectures and hands-on exercises. All participants will be trained in both semantic and nonverbal annotation schemes, but one or two course day during the week the participants will get the opportunity to focus on their preferred area of annotation. Tools: GATE, ELAN, ANVIL, and others.

Teachers: Teaching will be held by UCPH staff as well as by two invited international experts in the field. Suggested experts (to be confirmed): Paola Monachesi (Utrecht U), Martha Palmer (U Colorado, Michael Kipp (DFKI), Jens Alwood (Göteborg), and Brian MacWhinney (Carnegie Mellon U).

NEWDEVSCHOOL: CLARA Winter School on New Developments in Computational Linguistics (CUNI)

Goal: ESRs as well as Computational Linguistic masters students will be exposed to recent advances in Computational Linguistics.

Proposed topics and speakers: (a) Statistical parsing for language understanding by Prof. Charniak, head of BLLIP, laboratory for language processing at Computer Science, Brown University, Providence, RI, USA. He is interested in parsing (shallow and deep) for quite some time. He will give present recent hypotheses about contributions to parsing accuracy (b) Lexical resources for language understanding by Prof. Palmer, University of Colorado at Boulder, who is well-known for the creation of PropBank, a lexical resource linked to the world-famous Penn Treebank for verbal argument markup and sense disambiguation. She will present recent results in creating VerbNet and merging various lexical resources (PropBank, FrameNet and others) to a unified, high-quality lexical resource. Important issues regarding the sense granularity of such a resource will also be dicussed.

Time and format: The winter school is organized as a week-long event.

CAREERSCHOOL: Industrial Career Training Course: Product Planning for Next Generation Information Access Technology Solutions (UIB, COMPERIO and FFZG)

Goal: This training course is aimed at providing the next generation of scientists with the complementary skills that are necessary to move from theory to a marketable products and solutions. The course will include planning, market analysis, entrepreneurship, exploitation of research results, project management, proposal writing, communication, research ethics and IPR management.

Focus: The course will both demonstrate a complex industrial system based on multiple components and give an introduction to strategic product planning covering the span from idea to deployment. FFZG will present the case study of the system for semantic analysis of newswire texts that is being developed for Croatian News Agency (HINA). This case study demonstrates how a complex system can be build from simple existing modules such as lemmatization, POS/MSD tagging, named entity recognition and classification, document classification, keyword extraction; with perspective to widen the customer support with more advanced Knowledge Technology such as social network analysis, event detection, trend detection etc. The second part of the course starts with an overview of the principles of market and competitor analysis, followed by a session focused on project management. Basic concepts and benefits of agile development strategy (Scrum methodology) will be demonstrated by examples from industrial projects. The strategy of customer-driven innovation will be explained in detail, and methods will be taught to package experience and Best Practices in product development. The workshop will also discuss a selection of architectural issues, such as best practices in the implementation of clean, backward-compatible and standard-compliant programming interfaces (APIs); as well as the design of service-oriented architectures tailored to the next generation of information access technology solutions. Further modules will include complementary skills such as proposal writing, management, communication skills and ethics.

Organization, time and location: A four day course in Dubrovnik, Sep. 2011, organized by FFZG. The industrial associated partner COMPERIO will be the main responsible for the programme of the course; both partners will draw extensively on their industrial experience to make this workshop a cutting-edge event.

Time table of recruitment with starting dates and duration

Project	*Host*	*Researcher type*	*Duration*	*Starting date*
2A	MPI	ESR	36	12
2B	MPI	ESR	36	12
3B	UIB	ESR	36	6
3C	UCPH	ESR	36	11
3D	UPF	ESR	36	6
4A	NHH	ESR	30	4
4A	TILDE	ESR	30	6
5A	MPI	ER	24	1
5B	UCPH	ESR	36	11

6A	UTU	ESR	24	11
6B	TILDE	ESR	21	8
6C	TILDE	ESR	21	12
6C	CUNI	ESR	36	5
7A	UHEL	ER	24	1
7A	UHEL	ESR	24	4
7A	UTU	ESR	24	11
7B	UIB	ESR	36	6
7C	CUNI	ESR	36	13
7C	UTU	ESR	24	11

B.2.2 Planning of Visiting Scientists contribution

The following persons (or persons with similar expertise) will be visiting scientists contributing to the events described above.

- *Prof. Dr. Rita Temmerman* (visiting NHH for 3 months in WP4) is co-ordinator of Centrum voor Vaktaal en Communicatie (CVC) at Erasmushogeschool Brussel. Based on case studies on categorisation and naming in the life sciences (DNA technology) she has developed the sociocognitive terminology theory. This expertise brings a new theoretical perspective which will be highly valuable for the project and training course in terminological resources.
- *Prof. Dr. Sandra Kübler* (visiting UTU for 3 months in WP7) is affiliated with the Department of Linguistics, Indiana University. She provides outstanding expertise in treebanking which she will contribute to the workpackage on Parsing Technologies and Grammar Models and she will also contribute to the preparation of a thematic training course.
- *Prof. Dr. Martha Palmer* and/or *Dr. Paola Monachesi* (visiting UCPH for 2 months in WP3) will be invited to contribute to the research project and summer school related to Semantic and Nonverbal Corpus Annotation. Martha Palmer is affiliated with the University of Colorado; Paola Monachesi is affiliated with the University of Utrecht. Both are leading experts on the topic of annotation.

B.3 Impact

B.3.1 Research Indicators of Progress

B. 3.1.1 Research Activities

- General progress with research activities programmed at individual, participant team and network level. Possible problems encountered and nature/justification for adjustments, if any, to the original research work plan and/or timetable.
- Highlights of scientific achievements and recognitions (innovative developments, scientific/technological breakthrough, publications).
- Progress on cross interaction between academic and industrial partners.
- Visits of Senior Researchers from inside and/or outside the network.

- Individual and joint publications and conference presentations, directly related to the work undertaken within the project.

B. 3.1.2 Training Activities

- General progress with training programmed at individual, participant team and network level (Career development Plan, supervision, coaching or mentoring in place at each host institution).
- The rate of recruitment of ESR/ER for each participant and for the network as a whole (ratio person-months filled/offered) and time and duration of each individual appointment [Please note that these must be from 3 up to 36 months for ESR and between 3 and 24 months for ER. Short visits and secondments although part of the training are not counted as appointments, but as part of the networking activities.].
- The nature and justification for any deviation from the original plan (as refereed to table A3.1 of part C) or adjustments, if any, to the original research work plan and/or timetable.
- The number and place of the short visits/secondments undertaken or organised by each ESR and ER within the network (full participant and associated members including number of visits of the ESR and ER to their home scientific community).
- Organization of and participation in training events and network meetings (workshops, seminars, summer schools ...) and in international conferences (number, names, place date).
- Achievements regarding the acquisition of complementary skills (for example: project management, presentation skills, language courses, ethics, intellectual property rights, communication, entrepreneurship….).
- Level of satisfaction of the trainees (e.g. as expressed in response to questionnaire and their expectation to present their PhD thesis and when.).

B. 3.1.3 Management and impact

- Effectiveness of networking, communication and decision-making between partners (at all levels: Coordinator, team leaders, supervisor, ESRs and ERs), between the network and the Commission, and with the Industrial and/or other relevant stakeholders.
- Effectiveness of the recruitment strategy in terms of equal opportunities (including gender balance) and open competition at international level.
- Effectiveness of the "training events and conferences" open to external participants and integration in the training programme.
- Effective contribution of Visiting Scientists to the research training programme.
- Development of any specific planning and management tool(s) and databases management of intellectual property and commercialisation of network research output.(if applicable)
- Nature and justification for adjustments, if any, to the original training plan and/or timetable (e.g. opportunities for new collaborations regarding training activities).
- Cross-fertilization with other projects, esp. language technology and infrastructure projects such as CLARIN.

B.3.2 Dissemination and Impact

Plan for dissemination and IPR

Research results of CLARA will be normally be put in the public domain. Results will be published in the scholarly literature and presented at scientific meetings. Appropriate channels will be used for this purpose, e.g. the CL, MT, LLC journals, NEJLT, large European meetings such as EACL, regional ones such as NoDaLiDa and more specialized ones such as TLT. Furthermore, there will be a dissemination effort through the CLARIN project, using channels such as the CLARIN newsletter, website, and meetings. CLARA will put in place its own website with announcements of researcher positions, events and all research results. Finally, CLARA training events will also be open to external researchers so that these events will also serve a purpose in the dissemination and exploitation of the project. In particular, researchers in the CLARIN project will be targeted.

Europe's position in language technologies and contribution to standards

Due to its commitment to a highly multilingual society and its support for the maintenance of languages, Europe has a leading role in language resources and technology. Two of the largest programmes on documenting endangered languages have resulted in a huge treasure for linguistic analysis. Europe has leading roles in standardization iniatives such as ISO TC37/SC4 and TEI. Due to a number of projects and initiatives such as ISLE, INTERA, TELRI, DAM-LR and DOBES Europe has a leading role in building integrated and interoperable infrastructures that introduced federation and grid technology to the field of HLT. At the same time, Europe has a large number of outstanding experts in this sector, but their expertise has so far not been sufficientely exploited in researcher training. The CLARIN research infrastructure is confirming this strong position, will provide better accessibility of resources and technology to the European researchers, and will pass on the leading expertise to the next generation.

Creating a nucleus of young researchers across national boundaries

The combination of information technology and multilingual technologies and skills are a key to master the challenges of the semantic web. By further training specialists in this area, and making them aware of the industrial potential of the emerging technology, career prospects are drastically improved. The unique combination of experts in all relevant areas (HLT, standardization, federation/grid) is a great opportunity for Europe to educate a new generation of young experts who will overcome current limitations of the technology.

All partners will accept CLARA training as part of PhD programs and will recognize all training modules as such components. This, together with the high interaction between institutions in many countries in this ITN, will contribute to creating a European common high standard level of PhD training in this subject.

Transfer will be further enhanced by the significant number of open training events that the ITN will offer. Especially the 128 CLARIN members from 32 different countries will be able to make use of some the CLARA offers, which will contribute to the wide distribution of knowledge and the dissemination of methodologies across Europe. Finally, the visiting scientists will transfer their knowledge though their participation in network-wide events.

Creating synergies and long-term collaborations

All CLARA full partners and most associate partners, including industrial partners, are already planning very long-term collaboration at European level through CLARIN, which is currently in a three-year preparatory phase, to be followed by a ten-year construction phase. CLARA and CLARIN will mutually deepen this long-term collaboration, on the one hand through extensive knowledge and resource sharing in CLARIN, and on the other hand through CLARA training of the next generation of researchers for the CLARIN construction phase.

Coordination with international research programs

CLARIN has already close contacts with comparative initiatives in the US, Japan, Korea and Australia. Also institutions from countries such as South Africa, Argentina, Brazil etc are interested in a close collaboration with CLARIN. This will ensure that the methodologies chosen will be discussed at an international level. Due to the personal overlap between CLARA and CLARIN we are sure that the international perspective will also be considered in the CLARA work.

Promoting language diversity in a multilingual society

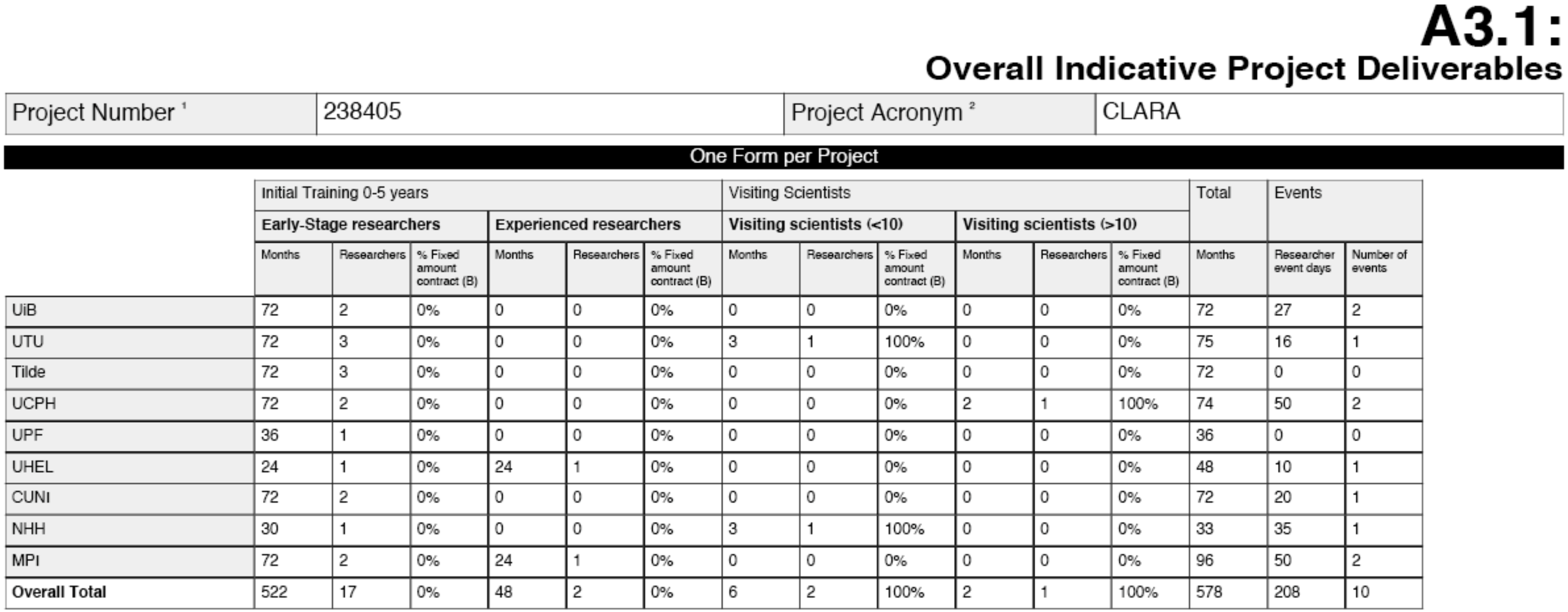
Europe is a multilingual society already facing enormous challenges with the 23 official languages in the EU (plus 2 more in the EEA), some 50 minority languages, and relations to numerous other language communities both within and outside of Europe. The European policy of "language diversity" needs to be supported by deeper linguistic research resulting in a better understanding of linguistic commonalities and differences, as well as in advanced language applications that will help to reduce the costs and improve the cross-cultural integration in a multilingual society as Europe.

B4. Ethical issues (not applicable)

B5. Gender aspects

The coordinator will stimulate the adoption of recruitment and employment policies in the consortium that stimulate fairness and equal opportunities, aim at a good gender balance and allow conciliation between private and working life. While legal conditions and trade union agreements in the contries and sectors concerned must be respected, the consortium will use whatever freedom it has to take into account these gender aspects as much as possible. Earlier experience indicates that this scientific field is equally attractive to female and male young researchers. In its recruitment policy, CLARA will present itself as attractive to both genders and will address itself to both genders. Gender-biased terms will be avoided and replaced by gender-neutral ones. Wherever possible, working conditions will be put in place that allow conciliation between private and working life. Researchers will be informed of employee rights such as maternity leave and of the availability of facilities such as kindergartens.

PART C: OVERALL INDICATIVE PROJECT DELIVERABLES



A3.1:
Overall Indicative Project Deliverables

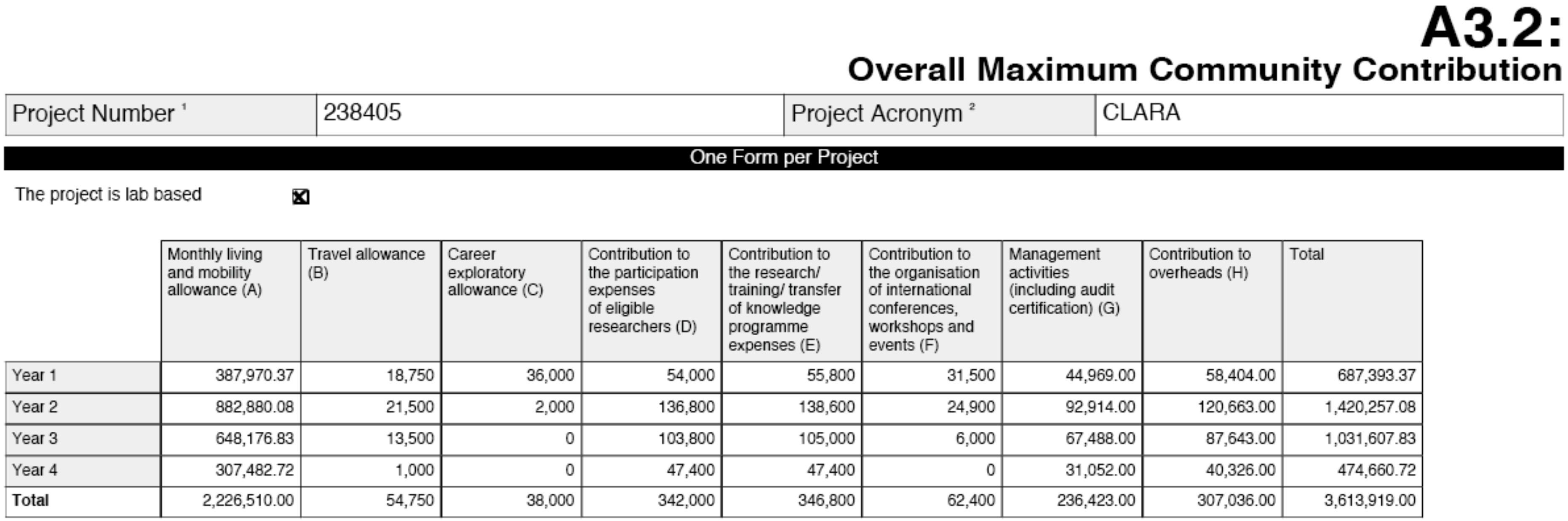
Project Number [1]	238405	Project Acronym [2]	CLARA

One Form per Project

	Initial Training 0-5 years						Visiting Scientists						Total	Events	
	Early-Stage researchers			Experienced researchers			Visiting scientists (<10)			Visiting scientists (>10)					
	Months	Researchers	% Fixed amount contract (B)	Months	Researchers	% Fixed amount contract (B)	Months	Researchers	% Fixed amount contract (B)	Months	Researchers	% Fixed amount contract (B)	Months	Researcher event days	Number of events
UiB	72	2	0%	0	0	0%	0	0	0%	0	0	0%	72	27	2
UTU	72	3	0%	0	0	0%	3	1	100%	0	0	0%	75	16	1
Tilde	72	3	0%	0	0	0%	0	0	0%	0	0	0%	72	0	0
UCPH	72	2	0%	0	0	0%	0	0	0%	2	1	100%	74	50	2
UPF	36	1	0%	0	0	0%	0	0	0%	0	0	0%	36	0	0
UHEL	24	1	0%	24	1	0%	0	0	0%	0	0	0%	48	10	1
CUNI	72	2	0%	0	0	0%	0	0	0%	0	0	0%	72	20	1
NHH	30	1	0%	0	0	0%	3	1	100%	0	0	0%	33	35	1
MPI	72	2	0%	24	1	0%	0	0	0%	0	0	0%	96	50	2
Overall Total	522	17	0%	48	2	0%	6	2	100%	2	1	100%	578	208	10

(from A3.1 of the GPFs)

PART D: OVERALL MAXIMUM COMMUNITY CONTRIBUTION



A3.2:

Overall Maximum Community Contribution

Project Number [1]	238405	Project Acronym [2]	CLARA

One Form per Project

The project is lab based ☒

	Monthly living and mobility allowance (A)	Travel allowance (B)	Career exploratory allowance (C)	Contribution to the participation expenses of eligible researchers (D)	Contribution to the research/ training/ transfer of knowledge programme expenses (E)	Contribution to the organisation of international conferences, workshops and events (F)	Management activities (including audit certification) (G)	Contribution to overheads (H)	Total
Year 1	387,970.37	18,750	36,000	54,000	55,800	31,500	44,969.00	58,404.00	687,393.37
Year 2	882,880.08	21,500	2,000	136,800	138,600	24,900	92,914.00	120,663.00	1,420,257.08
Year 3	648,176.83	13,500	0	103,800	105,000	6,000	67,488.00	87,643.00	1,031,607.83
Year 4	307,482.72	1,000	0	47,400	47,400	0	31,052.00	40,326.00	474,660.72
Total	2,226,510.00	54,750	38,000	342,000	346,800	62,400	236,423.00	307,036.00	3,613,919.00

(from A3.2 of the GPFs)

Appendix

Table 3.1: Reference rates for monthly living allowances (cost of living index 100)[25]

Researchers Categories	A (EUR/year)	B (EUR/year)
Early-stage researchers	**34 500**	**17 250**
Experienced researchers (4-10 years experience)	**53 000**	**26 500**
Experienced researchers (>10 years experience)	**79 500**	**39 750**

This amount represents an increase of roughly 1,9% of the 2007 Work programme, reflecting the average inflation in the EU during the intervening period as published by Eurostat.

Table 3.2. Travel allowances

Distance[1] (km)	Fixed-amount contribution (EUR)
< 500	250
500 – 1 000	500
1 000 – 1 500	750
1 500 – 2 500	1 000
2 500 – 5 000	1 500
5 000 – 10 000	2 000
>10 000	2 500

For researchers eligible to receive travel allowances, the allowance is based on the direct distance (in a straight line) between the place of origin and the host institution of the researcher, calculated on the basis of one payment for every period of 12 months or less, when the first period or the last one is less than 12 months. Only one travel allowance shall be paid per period of 12 months, independently of possible interruptions or stays with different partners.

[25] Rates for individual countries are obtained by applying to these rates the correction factors for cost of living, as referred in Table 3.3