Attachment 1 – Proposal of the Project of large infrastructure approved by the Government

LINDAT/CLARIN Project
Establishing and operating the Czech node of pan-European infrastructure for research

The applicant

The Charles University in Prague is the applicant, however, the Institute of Formal and Applied Linguistics of Faculty of Mathematics and Physics, Charles University, will administer the funds. The Charles University in Prague will ensure in accordance to bilateral contracts a transfer of the funds necessary for operating the Clarin central node and for performing the national tasks at the workplaces which will contribute to the project based on the approved budget. In addition to it, the Charles University in Prague will provide for (if it is not otherwise decided so far) the transfer of scheduled resources to future Clarin-ERIC associations.

A person responsible for organisational, technical and personnel management of the project (principal investigator): Prof. RNDr. Jan Hajič, Dr. (hajic@ufal.mff.cuni.cz), The Head of the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Malostranské nám. 25, 11800 Praha 1 (phone: +420 607 209 212).

The Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague (http://ufal.mff.cuni.cz) is a leading national centre in the field of computer processing of natural language which has been developed in the framework of research centre programmes and research tasks, holding an outstanding international position in existing pan-European projects. The planned project of the research infrastructure directly follows the CLARIN (FP7-RI-2122230) project. The applicant's workplace is a national coordinator of the above project.

- ***Description of the research structure***

1.1. Conception

LINDAT-Clarin is designed as a Czech "node" of Clarin, the international network (Common Language Resources and Technology Infrastructure, FP7-RI-2122230, so far 2008-2010) which has been extended to the "T4ME Net" Project (Technologies for the Multilingual European Information Society, NoE, 2011-2014, FP7-ICT-4-249119) for free sharing the language data and basic technologies between institutions and individuals both in science and in research. The Clarin Project is more focussed on humanities, while T4ME is more general and relates prevailingly to language technologies and applications.

The LINDAT-Clarin Centre will therefore link both of these areas and especially their annotations (i. e. formal manual, semiautomatic and automatic language analysis) in the Czech language environment and will specialise in collecting language data and particularly their annotations. Collecting and annotating will be performed in a scale, quality and technological preparation (specifications, schemes and formats), which will be directly applicable both in humanities (linguistic and interdisciplinary research where language plays

a crucial role), and in research and development in language technologies using modern statistical and hybrid methods.

The work content of the Centre and its results penetrate to a number of branches, e. g. general linguistics and linguistics specializing in specific languages, particularly Czech language, translating, lexicography, sociolinguistics, and partly also related disciplines (psychology, sociology, library science, neuroscience, cognitive science) with an outstanding overlap to information technology (computer science, computational linguistics), mathematics (statistics and probability) and electrical engineering (processing the acoustic signal).

In the view of national priorities of the applied research (in accordance with the document Národní politika výzkumu, vývoje a inovací České republiky na léta 2009 – 2015, Hlava VI)/National policy of research, development and innovations of the Czech Republic for 2009-2015, Chapter VI) the proposed centre falls under the priorities 6 (Information society) and 8 (Priorities of the development of the Czech Republic).

1.2. Current state

Language resources and technologies for their processing in the respective European countries (like in the U.S.A. and Asia) have already existed, however, the present centralized distributional agencies (Linguistic Data Consortium in the U.S.A. and European Language Resources Association in Europe) do not meet the up-to-date requirements for simple, unbureaucratic and above all free access to language data for the majority of scientific, research, and development communities.

As a result, the data distribution is fragmented and uncoordinated with all negative impacts (incompatible formats and resulting difficult software applicability for their processing, a lot of different conditions for awarding licences, often an impossibility to gain a direct access to the data itself followed by a necessity to use lots of various searching engines etc.) The data in the Czech Republic are collected and annotated mostly at four workplaces that should contribute to the activities of the LINDAT-Clarin Centre, namely: the Charles University in Prague, the Masaryk's University in Brno and the University of West Bohemia in Pilsen together with the Institute Of the Czech Language of the Academy of Sciences of the Czech Republic in Prague. These workplaces presently constitute the Centre Of Computational Linguistics (CKL – a project of the Ministry of Education, Youth and Sports LC536), which is, however, a scientific workplace producing the language resources rather marginally and is going to be dissolved after 2010.

Clarin, as a project of the 7th Framework Programme (ESFRI), is focussed on eliminating the above shortcomings in the European framework in favour of humanities and social sciences, particularly research in all linguistic disciplines. The T4ME Net Project (to be launched in March 2010) is oriented in a similar way, but its primary aim is to serve to the interlinguistics community (linguistic, information science, statistics), mostly in language technologies, so called computational linguistics, which affects the related applications.

1.3. Subject structure of the project

In accordance to the project concept and its interconnections to the Clarin and T4ME Projects the project will be structured as follows:

1. Main part: the node of Clarin, the distributed European network

2. National part: Czech and multilingual language resources: collection and production of language corpuses and database
3. Coordination part: coordination in legal, technological, educational, and external relations fields

The organisation structure of the Project (see Chapter □) will be created and the Project budget will be segmented (Chapter □) based on this subject structure.

1.3.1. The node of Clarin, the distributed European network

As it is clear from the nature of the project and its linkage to above all ESFRI and Clarin project, this part of the project is principal for the storage and providing data and services which are the core of the Clarin Project. The LINDAT-Clarin Centre will as a distributed node provide for the technological background for A- type node (based on the definition of the Clarin project in the phase of preparation), i. e. the highest model of such distributional node. The A-type node provides for the storage of all language data (including accepting new data and their integration in the system including allocation unique persistent identifiers), authorised access using pan-European federation of entities, allocating identities for the users of the system, and local web services (fully) and distributed ones (mediated) for processing language data, access to them and their passing to other Clarin nodes. This project part will be executed by the applicant through the Faculty of Mathematics and Physics, Charles University in Prague.

Making decisions on directing this part (unlike the national part, see 1.3.2) will proceed in the structure of the Clarin project. The Czech part will contribute to this decision according to the rules of the Clarin project, as well through the Faculty of Mathematics and Physics, Charles University in Prague, which has been entrusted with coordinating the national activities of Clarin in the Czech Republic.

1.3.2. Czech language resources: collecting and producing language corpuses and databases

Whereas language data, which are to be stored and shared by the distributed node of the Clarin network, exist (and have been constantly produced) at a lot of workplaces abroad, it is necessary to actively support collecting and producing language corpuses under the conditions in the Czech Republic. From clear reasons, language data abroad originate in the corresponding local area (German in Germany, Dutch in Holland, French and Flemish in Belgium etc.) therefore the Czech language data should be acquired in the Czech Republic.

Whereas also other workplaces in the Czech Republic are involved in collecting data and LINDAT-Clarin activities will only complement the missing fields, especially spoken data, parallel data (with other languages) in a broad scope, and combined data (see the below definition) in producing annotated language corpuses, which form the basis for further research and development both in humanities and in technological and application area, the LINDAT-Clarin Centre in the Czech Republic will be a unique project.

Collection of language data means:

collecting written (possibly spoken) language data from publicly available sources (internet, open sources); these data can be combined with other modalities e.g. video, sign language recorded in symbols etc.;

contractual collecting written or spoken language data (with a possible visual segment) from publishers and owners of such data;

recording speech data in a predefined environment (laboratory, adapted to the application, field work etc.);

electronic data from the sensors when the subjects are exposed to the natural language in any form (video, haptics, electrical or magnetic scanning of the reactions etc.)

Producing language corpuses and databases means such processing the selected data that enriches them for using in humanities and technical sciences in the aspects as follows:

so called "clearing", unification and normalisation of data to standard data formats which can be used for further processing or distribution;

alignment for parallel language data (written, spoken, visual or other) with the aim to establish explicit relations between languages or various data modalities, by means of manual or automatic methods, the result of which are synchronisation signs of content or time character;

extraction of language databases from language data while under the databases we mean above all dictionaries (phonetic, morphological, syntactical, and semantic …, possibly related to ontology)

annotating language corpuses; annotation means manual or (semi)automatic analysis of language data (with an output in electronic format), which is necessary for further usage of data in humanist and technological fields, according to the world standards and new approaches with a view to the characteristic features of the Czech language.

This part of the project is aimed at the preparation of language data (language corpuses and databases, as the case may be including annotation) to be published in the shortest possible period through the Clarin node (see 1.3.1) and other distributional channels. They should be at disposal to the general public through a simple licensing and authorization policy without any technological and legal obstacles from the institutional level to the level of the individual investigator or student.

All of four Czech partnership workplaces will contribute to this part of the project according to their orientation and specialization. A common approach is expected in the respective cases when producing particularly demanding language corpuses and databases as to the processing or technological aspects.

1.3.3. Coordination in the legal, technological, educational, and external relationships areas

The collection, production, annotation and distribution of language data requires based on existing experience a lot of successive activities which are advantageously concentrated externally to the linguistic or technological work.

For collecting data on one hand and their distribution on the other hand it is necessary to pay a consistent attention to an appropriate active and passive licensing policy with the aim to make the produced language data as simple as possible for the end users, and in addition to it,

provide the users with a maximum legal certainty that the data which they acquire through the Clarin distributional centre are not limited in any way in the respect of scientific application. The situation now in the EU is simpler than several years ago due to the changes in the copyright laws and regulations (in the Czech Republic see law 121/200 Coll., as amended), nevertheless it is not optimal for scientific usage language data therefore it is necessary to back it up appropriately.

The technological aim is to consolidate software tools for collecting, clearing, alignment and (manual) annotation of data so that the interoperability minimally at the level of data formats (XML and their variations and restrictions for language area) could be ensured. The identical aim is necessary to be reached for storage and distribution of data and web services in the Clarin network (as it has been decided in accordance to the Clarin coordinating centre).

Education and training of professionals in the field of language, language-technological and organisational fields in all working areas of the LINDAT-Clarin Centre at the level of follow-up master's programmes, the Ph.D. level and further training of the experts from universities, Academy of Science and applied science. The aim is a well-educated group of workers who will further effectively operate the LINDAT-Clarin Centre in the humanistic and technological research and in practise regarding language data and technologies.

The LINDAT-Clarin Centre will as well broadly promote a usage of language technologies and demonstrate their possibilities for developing knowledge society through holding seminars intended for the respective target groups (from the scientific to the managerial ones) by contributing to organising pan-European activities and events in the research-linguistic field (especially in the cooperation with the T4ME Net project) and by participating the workers of the Centre in the activities of such type.

The above activities will be coordinated both with the Clarin project (or with its head office), and the T4ME Net project at the "European" side and also from the view of national participations (University of West Bohemia, Masaryk's University, Institute Of the Czech Language of the Academy of Science of the Czech Republic) through participation of the representatives of Faculty of Mathematics and Physics, Charles University in these projects and internal organisational structure of the LINDAT-Clarin project (se Chapter □).

- ***Benefits for the research and development in Europe and the Czech Republic***

The Centre is going to continue in the international cooperation in data producing and distributing. Nowadays, 171 institutions from 27 member and 5 associated EU countries (at the level of universities, academic institutions tec.) work together in the Clarin project. At the start, 17 partners will participate in the T4ME project financed from EU funds that will establish a consortium of users with approximately 200-300 institutional members from EU and associated countries. The Charles University (represented here by the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics) is a partner from the Czech Republic in both projects.

Potential project partners were and still have been involved in another projects both in the Czech Republic and in Europe, in which there (so far in an uncoordinated way) originate language data, which are distributed in different ways, e. g. EuromatrixPlus and Faust (Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague - language data pro automatic translation), Companions (Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics, Charles University in Prague, Department of Cybernetics FAS University of West Bohemia – speech annotated data pro the development of dialogue systems), KYOTO (Masaryk's University Brno, lexical data). In addition to it, the Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics, Charles University is engaged in NoE FlareNet, an existing smaller project, where the standards and specifications of the common source for language resources have been created.

The Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics, Charles University also cooperates with U.S.A. and some Asian workplaces in the projects of language data annotations: NSF PIRE project (annotation for advanced understanding of spoken language, machine translation: Johns Hopkins University, MD, USA, Brown University, RI, USA), the projects of syntax annotations and a valency (University of Colorado, CO, Brandeis University, MA, IIIT Hyderabad, Indie), the project of language data annotation for the discourse research (University of Pennsylvania, PA, USA). Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics, Charles University also administers a part of the data and access point to 116 thousands of memories records of victims having survived the holocaust ("Malach Centrum", Contract with the University of Southern California, CA, USA), which are an outstanding and still unsurpassed source as to the size of spoken data in many languages.

The Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics, Charles University has long-term experience together with a Department of Cybernetics of FAS University of West Bohemia in Pilsen with publishing and "classical" distributions of language data through the centre of LDC (USA) and ELRA/ELDA (France/EU).

The relevance of the proposed infrastructure can be summarized as follows:

1. It will form a national reference source of language data for publicly available, easy and legally feasible scientific, research and application- development usage;
2. it will enable a wide approach to the professional community and users to verified expertise;
3. it will enable a wide approach to the professional community and users to computer tools, technologies and service which have already been developed;

4. it will enable a wide approach to the professional community and users not only to monolingual (Czech) sources, but also to multilingual sources and corresponding technologies;
5. it will form a significant "added international value" to the national, in this case, Czech, initiatives, enabling linking centres throughout Europe;
6. it will provide a powerful potential for innovations;
7. it will strengthen an interest in an international language as a part of international culture and national heritage;
8. last but not least it will greatly contribute to the streamlining of the educational process (teaching languages, language technologies, big data and their processing)

There is a key question, who and how is going to use the LINDAT-Clarin Centre which has been established. Based on previous experience we suppose that the users will come from the categories as follows:

1. in the scientific field: all academic workplaces (both at universities and the Academy of Sciences of the Czech Republic) participating in scientific processing, teaching and computer processing); also foreign workplaces are supposed (workplaces of the same or similar type) in the framework of Clarin and T4ME Net projects), students involved in scientific work (master's and Ph.D. studies, foreign students in the Czech Republic);
2. in the field of application development and innovations: all workplaces and industrial organisations working with information systems (librarianship, documentation centres etc.), translations services, documents and information searching, terminological databases, supporting tools for creating texts (spell checkers), documentation in the field of history, sociological and psychological research and applications of a similar kind, in which texts and text records are used; there is a high probability of usage from abroad (localisation projects of foreign companies, companies developing tools supporting automated translations etc. – see the above list);
3. in the field of education: application in educational systems of all levels in language teaching, teaching IT, possibly in other subjects.

- ***Project goals***

3.1. Introduction

The EU/ESFRI Clarin project, like the future T4ME Net and LINDAT-Clarin Centre designed with the same idea, is going to gradually eliminate the obstacles to a free access to language data, and enable distributed, however unified providing language data and related technologies. The ambitions of Clarin and to a great extent also of T4ME (with the exception of evaluation data for complete testing the tools for processing natural language) lie especially in unifying technologies and distribution – their format has been left to a capacity (and financial support) in the respective states, since these are often national languages. The collection and annotation (i.e. data production) is therefore an inseparable part of the proposed LINDAT-Clarin Centre.

Annotating data in an appropriate extent is necessary for applying the results of such research in practice (spell checker, automatic translation, information extraction from the text, understanding the text, dialogue systems etc.), because the development of software tools is based on statistical methods which require a lot of mainly linguistically interpreted

(annotated) data. These data are required also for including Czech language in the localization programs of text products of large companies which work with these methods.

The existence of annotated data in the relevant language is often a decisive factor for including the institution in large research projects in the EU (Charles University was in past awarded several projects of such kind also thanks to its ownership of annotated corpuses – e. g. Euromatrix, EuromatrixPlus, Companions - with the University of West Bohemia, Faust, and of course the Clarin and T4ME Net projects as well).

Annotation of the data in an appropriate extent is very demanding as to time and capacity; it is mainly manual work of highly specialized linguistic professionals, i e. Ph.D. students (exceptionally follow-up master's) of linguistic, and especially interdisciplinary subjects combined with informatics, with a high proportion of technical contribution by experts in information science. Data collection is a demanding work as well which requires experts of various professions from information technology through organisation tasks and legal background to the training and education of users.

The unique distribution will provide for more general publicity and usage of data and therefore a better utilization of invested funds.

It is a non negligible aim of the project to prepare further scientific generation able to work with language data, create and use them properly both in national and international context, and work together both within and outside EU at future projects using language technologies.

3.2. Interim goals of the project

Interim project goals are based on two main project objectives described in the previous paragraph: – collect language data for Czech language (a parallel data) and provide an access to them. Speaking about continuous interim goals of the project, they contain preparing a young scientific generation in the area of language data both in humanistic and technological branches.

3.2.1. Constructing Clarin, the distribution node

The construction of the distributional node of "A" type in Clarin methodology is a critical condition of operating the LINDAT-Clarin Centre in the framework of European research area and Clarin (and T4ME Net) project. This node is going to be built gradually in 2010s (preparation, completing the specifications) and 2011-2013 (construction phase, concentrating technologies, gradual collecting data and launching testing operation). The full operation is expected in 2014.

The construction will have an investment part (in the sense of procuring computational and network technology at the Charles University in Prague and other workplaces) and a software part concerning launching a worldwide unique data identification and an authentication of users valid at least throughout Europe by means of identities federation (EduID in the Czech Republic:), establishing a negotiated (in Clarin framework) repository for storing their own and "foreign" data, introducing a unique description of metadata, introducing a unique API for web services concerning language data processing and providing a legal framework (sample contracts for open licences for data providers, licences for users).

3.2.2. National infrastructure for collecting and producing language data

The respective workplaces of the suggested LINDAT-Clarin Centre have been already cooperating in the areas of data collecting and annotating in the respective cases ("Companions" project as a part of the 6th Framework Programme etc.). Within the proposed LINDAT-Clarin Centre the tools for building and annotating language data will be handed over so that the highest affectivity and availability can be reached also outside the partnership organisations.

The tools for collecting data from open sources, for basic processing (data "clearing", tokenisation, segmentation) and for basic language automatic pre-processing will be mutually freely accessible. Provided that it is effective, they will be unified (if similar, however, not identical types of tools, which are satisfactory for everybody, are presently used).

For data distribution and services regarding language data processing where it will be more efficient also within the Czech Republic, Clarin node will be used (see 3.2.1).

3.2.3. Data collection

Data collection will be performed depending on the type of the data. Collecting the Czech and text parallel data from the open sources (i. e. mainly from the Internet) is a question of a suitable technology which is going to be shared by the respective Czech workplaces.

It is necessary to get the speech data from the subjects, i. e. people who will read suitable passages for the specific task. Recording dialogues and sensory responses of people to language incentives will proceed in a similar way. In all cases an approval of the relevant subject in accordance to the law on the protection of personnel data is required; however, the nature of collecting data does not require special provisions regarding ethic standards.

Initial data processing and their storing with unique metadata in the format suitable for further processing, especially manual transcription and annotation (see 3.2.4) is a part of data collection.

3.2.4. Data annotation

Linguistic annotation of data, which is crucial for evaluating the data especially for their future application utilization, is a most demanding operation when preparing them with respect to both time and financial resources. The basic annotation method consists of four phases:

- selecting the data (from collected and prepared data);
- preparation of the annotation tool for direct annotation in electronic form;
- actual annotation (including organisational and administrative background);
- checking the annotation quality, preparation of the data for storing and making them publicly available;

As to the amount of work attributable to the respective phases, the annotation forms about a half of work. The quality control also depends on the complexity of annotation – it consist mostly in comparing double or triple annotation and correcting inconsistencies between

annotators in data; various automatic or semiautomatic methods of checking the formats and contents of annotation can be applied.

3.2.5. Preparation of the data to publication, making them available and data distribution

The goal of the LINDAT-Clarin project and the actual European Clarin and T4ME Net projects is to make available (to a maximum extent and in an easiest way) language data and methods of their processing to the research community in the Czech Republic. Therefore the data can be published – in the new Clarin it means, that data will be after their production and annotation available in the Clarin network nodes as soon as possible.

For such purpose it is required to describe the produced data from the point of view of the future users (annotation manuals, their translations to relevant languages), and store the data correspondingly in the repositories of the Clarin network.

3.2.6. Integration of software tools in provided web services

It is necessary to create software for an approach to language data including searching and statistical modules (if the data are not made available by simple providing them as a data copy) and implement them as a web service in the Clarin network. In the respect to a variability of the individual data files and databases it is often necessary to produce specialized software which will enable users e. g. from humanities make use of the data in a full and effective way.

3.2.7. Preparing scientific workers in the field of language data

Preparing scientific workers in the field of language data forms an inseparable part of the LINDAT-Clarin Centre both from the linguistic (theoretical and applied) and of the information technology aspects. All four workplaces have experts in these fields; there will be also used (if students involved in the work at the centres are interested) accredited master's and Ph.D. programs because both language data and methods of their collection are a suitable material for dissertations and Ph.D. theses.

The LINDAT-Clarin will as the case may be provide education for student both in the EU and outside EU according to valid regulations and directives making use of EU training programmes or in a direct way.

3.2.8. Organisation of scientific seminars and conferences in the field of language data

For the experts (of all the relevant fields) the LINDAT-Clarin Centre will hold seminars or contribute to the organisation of conferences in search and natural language processing specialisation where the cooperating organisations have a long term tradition (Text Conference, Speech and Dialogue in the Czech Republic, conferences and workshops of the Association for Computational Linguistics).

3.2.9. Organisation of seminars and trainings for service users

The service users of the Clarin node will be trained in the methods of using language data made available through this network both in the Czech Republic and in cooperation with the Clarin head office abroad.

3.3. Time schedule for 2010-2015

The time schedule can be generally divided into three parts:

- 2010: preparation of the Centre
- 2011-2013: construction phase, in cooperation with Clarin (Clarin-ERIC, if its is established)
- 2014-2015(-2020 and further): operation (operational phase) / Clarin-ERIC node.

Presently, the language resources and relevant programme tools are stored at the workplaces where they have been produced and they are also made available or distributed by these workplaces (www, physical media). They are exceptionally distributed by the two existing international centres (LDC - USA, ELRA - France). Creating the sources is performed especially at the workplaces of Charles University and the University of West Bohemia, Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics, Charles University has build a smaller storage of several TBs of the disk space) and a computational capacity of approximately 200 CPU units for basic language data processing.

1. Building storage and distributional centre ("A" in the intentions of the Clarin project and a similar centre in the framework of T4ME project) is supposed to be formed in two phases: (i) building a technological infrastructure (storage, networks, connection) to 2013 (construction phase), and then (ii) its operation and possibly extending the capacity (operational phase, continuously). This distributional centre is going to be at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University.
2. Data collection, archiving, making them available: this activity is performed in the Czech Republic concurrently at several workplaces (Charles University, Masaryk's University a University of West Bohemia). All these activities should be transferred to a common platform (integrating relevant data collections under the distributional scheme of the centre referred in (1) until the end 2013. Operational phase continuously – partly simultaneously with the construction phase, fully from 2014.
3. Specification and production of annotated data, software tools for data processing: transferring these activities and partly several "Research plans" from the contemporary Centre of computational linguistics and after their finishing, providing organisational background, connecting to distributions referred in (1) till the end of 2013. Operational phase (including providing tools and capacities for the preparation of the data) continuously from 2014.

- ***Present financing of infrastructure and integrating the Czech Republic***

Clarin: budget for the Czech Republic (administered by the Charles University): 90,000 euros plus co-financing by the Ministry of Education, Youth and Sports of the participation in the framework programmes – at the level of 33 % total costs of the Czech participation in the project (26,000 euros), 2008-2010.

Flarenet: 6th Framework Programme, total budget for the Charles University: 9,000 euros, without co-financing, 2007-2010.

The parts of EU projects intended for producing language data: approximately 660,000 euros plus a part of the contribution to the Czech Republic for the projects of the 7th Framework Programme: approximately 100,000 euros for various projects in 2006-2013 (scheduled).

Data preparation, project in the U.S.A.: after recalculation about 150,000 euros (source: NSF, DARPA, 1998-2009)

National projects ("250 million" center, centers of type LN and LC, project MSM of the MEYS, Grant Agency of the Czech Republic, Grant Agency of the Charles University, other - Ministry of Education, Youth and Sports) within which language data are collected or annotated: calculating approximately 1,800,000 euros as a total in 1996-2010 (rough estimation) for actual data collection, annotation and data distribution.

Structural funds cannot be used for the collection and acquisition of the data.

So far, total (1996-2009): 2,835 thousands euros, of it from the resources in the Czech Republic: 1,926 thousands euros,
(in CZK: 70 million Czech crowns, of it from the resources in the Czech Republic: CZK 48 million)

The above resources have been intended for data collecting and annotating therefore they are administered not as *investment costs*, but as *running costs* (investment costs are of a lower order, only for data storage and their updating - depreciation). In terms of the overall character of language infrastructure, however, the costs for collecting and especially annotating and processing language data always prevail.

- ***Project duration and the respective phases***

The project is scheduled for 10 years (with a possibility to be continued), i. e. for 2010-2020. In 2010 preparatory and organisational works will be carried out from July 2010 (expected start of providing the support) (concurrently with finishing the preparatory phase of the Clarin project under the 7th Framework Programme), while full integration to the Clarin network is expected from 2011, like the start of works at a national deployment and coordination of associated activities.

There are three project phases: the preparatory phase (2010), the construction phase (2011-2013) and operational phase (2014-2020 and further).

- ***Indicative framework project budget***

Estimation of costs in thousands of CZK:

Preparatory phase in 2010 (initial investment, legal and organisational preparation) 15,000
Construction phase, 2011-2013: 60,000

Operational phase, 2014-2015: 80,000
Estimation for 2016-2020 (operational phase, extending, new technologies): 210,000

Note: According to the data from ESFRI the construction phases will be financed from the national funds, a small direct contribution from EU cannot be expected until operational phase.

Of it the requirement for financing in the Czech Republic:

Preparatory phase (initial investment, legal and organisational preparation, mechanism of supplementary funding of framework programmes)
Construction phase, 2011-2013: 60,000
Operational phase, 2014-2015: 60,000
Estimation of costs for 2016-2020 (operational phase, enlargement, new technologies): 150,000

Detailed substantiation and a breakdown for the individual budget categories are specified in the part of this proposal.

- ***Indicative detailed budget for construction project phase (2011-2013) and first two years of operational phase (2014 – 2015)***

Breakdown according to the structure of the planned LINDAT Centre (2010-2015):

In thousands of CZK		Year						Total
		2010	2011	2012	2013	2014	2015	2011-15
1 Data distribution	**Total**	**6,777**	**4,405**	**3,447**	**3,519**	**3,520**	**3,521**	**25,189**
	Inv.	*4,800*	*1,000*	*0*	*0*	*0*	*0*	
	Personnel incl. social security and health insurance (SHI)	*642*	*2070*	*2085*	*2135*	*2,136*	*2,137*	
	Other running costs	*744*	*305*	*320*	*320*	*320*	*320*	
	Indirect	*591*	*1,030*	*1,042*	*1,064*	*1,064*	*1,064*	
2 Collection, annotation	**Total**	**6,993**	**11,724**	**12,560**	**12,575**	**12,575**	**12,575**	**69,002**
	Inv.	*5,050*	*900*	*0*	*0*	*0*	*0*	
	Personnel incl. SHI	*340*	*6,650*	*7,800*	*7,800*	*7,800*	*7 800*	
	Other running costs	*1,030*	*950*	*1,020*	*1,030*	*1,030*	*1,030*	
	Indirect	*573*	*3,224*	*3,740*	*3,745*	*3,745*	*3,745*	
3 Coordination	**Total**	**1,123**	**3,802**	**3,802**	**3,802**	**3,802**	**3,802**	**20,133**
	Personnel incl. SHI	*0*	*0*	*0*	*0*	*0*	*0*	

	Other running costs	*680*	*1,745*	*1,745*	*1,745*	*1,745*	*1,745*	
	Indirect	*100*	*180*	*180*	*180*	*180*	*180*	
	Personnel incl. SHI	*343*	*837*	*837*	*837*	*837*	*837*	
	ERIC contrib.	*0*	*1,040*	*1,040*	*1,040*	*1,040*	*1,040*	
Total		**14,893**	**19,931**	**19,809**	**19,896**	**19,897**	**19,898**	**114,324**

Scheduled costs for language infrastructure –the LINDAT Centre, in thousands CZK

Substantiation of the budget

A commentary to the respective cost items (the brackets include the reference to the section of the proposal describing the content of work and funded activities – see Proposal, Chapt. 1.3):

General and common scenario

"Personnel costs" are always stated including mandatory payments to social security and health insurance (SHI) and to the Social Fund of the Charles University and participating institutions (presently it is 36%).

Indirect costs (overheads) are based on internal directives and calculations of the respective institutions and correspond to the overheads asserted towards the EU or to an average of the internal regulation and indirect costs applied within the EU. Indirect costs are not calculated for the ERIC-Clarin project. The respective coefficients are 0.44 (Charles University), 0.4 (University of West Bohemia), 0.4 (Masaryk's University) and 0.4 (Institute Of the Czech Language).

In the area of personnel costs (see below for the respective areas of work in LINDAT-Clarin) all four workplace presently have workers with appropriate qualifications for filling up the scheduled 12-14 permanent working positions, both from the professional, and the managerial point of view, even in the long-term horizon.

Distribution (of data and services) (1.3.1)

The investment costs: in this part of the project (providing technical background of the Clarin node by at the Charles University in Prague) is necessary to build also back-up network infrastructure for ensuring continuous operation and significantly strengthen computational and storing capacity for providing web services connected to operating the "A type" Clarin node. The main investment will be made in 2010 and finished in 2011; no more investing is planned for the next two years. Other enlargement will be effected in 2014 or later according to the actual utilisation of the Centre in 2014 or later.

The personnel costs are planned also mainly at the Charles university in Prague which will provide for the operation of the node of the Clarin network. The budget follows from a necessity of preparation and investments in 2010 (half-time work of a technician) and 2 positions (technological background (1), software – development, maintenance and support (2 working positions)) at the time of construction (and testing operation). A scientific coordinator of LINDAT Centre is included in personnel costs for a half-time work load (at the Charles University). Personnel costs at the other three institutions in this part concern

the data and software installation and maintenance for web services supplied by these institutions as a part of the Clarin Centre node.

Other running costs relate to the procurement of minor possessions and software licences which will be prescribed uniformly from the head office of the Clarin Centre and which do not exist in the open source environment, particularly repository software and software for authentication and creating identifiers for data files (paid service: will be executed by the Clarin head office or separately by the individual nodes of the Clarin network).

Indirect costs: see the general part of the commentary above.

Data collection and annotation (1.3.2)

Investment costs: it is necessary to increase computational power at all workplaces of LINDAT-Clarin Centre (in the framework of the existing computational clusters), especially the capacity and backing-up of data storage. The strengthened computational capacity will be used for data processing before, at and after their annotation and for preparing the data for publication. These investments will be acquired in the first two years.

Personnel costs form a most significant part when collecting and recording data. The personnel costs can be generally divided into three categories:

- costs for permanently employed workers (linguistic and information technology experts for linguistic, organisational and software support for data annotation);
- costs for annotation works, especially those of linguistic character (students and to a smaller extent experts in the field of linguistics; mostly based on the agreements to perform the job for terms corresponding to the duration of the specific project;
- costs for subjects when collecting the data (records of acoustic and other sensory data for which it is necessary to hire external workers, mostly for a limited period of about several hours.) The contract is in most cases the agreement to perform the job.

It is expected that there will be more (5-8) annotating projects carried out simultaneously each time, mostly at the Charles University in Prague, where some long-term annotation projects (e.g. Prague dependency corpus: lexical and structural semantic, discussion annotations, conferences of various types) will be finished in the LINDAT-Clarin Centre and further projects will be executed (collecting and annotating data for dialogue systems, collecting large corpuses including parallel ones in the order of milliards of words, data for experiments in cognitive linguistic, work at dictionaries with full language coverage both in Czech and English – morphological, syntactical and semantic ones). Projects of collecting and annotating search data will take part in other workplaces of the Centre (mostly at the University of West Bohemia in Pilsen), producing very large (milliard-) corpuses and dictionaries of various types including the specialised ones like Czech WordNet and following semantic networks and ontology (Faculty of Information Technology, Masaryk's University in Brno) and Czech lexical database based on it as a supporting material for new Czech dictionaries, making accessible new historical databases and their maintenance and extending (the Institute Of the Czech Language of the Academy of Sciences of the Czech Republic in Prague) and creation of corpuses and dictionaries of various types including the specialized ones (at Masaryk's University in Brno and in the Institute Of the Czech Language of the Academy of Sciences of the Czech Republic in Prague: lexical databases, making accessible historical databases, their maintenance and extending).

Forecasted numbers of workers at the respective workplaces (the base of calculation of personnel costs in the above 3 groups, only for the purpose of collecting and annotating Czech and parallel language data):

Workplace:	Charles University in Prague	Faculty of Information Technology, Masaryk's University, Brno	Faculty of Applied Sciences (FAS), University of West Bohemia, Pilsen	Institute Of the Czech Language, Academy of Sciences of the Czech Republic
Permanent employees	3-4	1	1	1-2
Annotations, cont.	15-20	5-6	6-8	3-4
Subjects at total	40-60	0	50-70	0

Numbers of permanent workers at the moment of starting the construction phase (2011)

Coordination (1.3.3)

Within this project branch a coordination with the European management of the Clarin project will proceed (Clarin-ERIC) at the construction and later in the operational phase. The fee to the central node (The Netherlands, the letter of Dr. Jeannette Ridder-Numan to ing. J. Marek from 09.12.2009) is planned to be 40,000 euros (CZK 1.1 million according to the present exchange rate).

With respect to the preparatory phase (2008-2010), these costs will not be considered until 2011.

Other costs are labour costs for a coordinator / professional guarantor (half-time work load) and an administrator and financial coordinator (half time), and travel expenses for the meeting of Coordinating Committee and Clarin Scientific and Advisory Committees for other LINDAT-Clarin workers.

Breakdown of costs for the respective institutions of the Centre

The breakdown is based on the overall table and expertise of the respective workplaces. Only total costs in the respective monitored costs categories are specified herein.

Charles University in Prague (Faculty of Mathematics and Physics, The Institute of Formal and Applied Linguistics)

Charles University v Prague	Year			
Type	2010	2011	2012	2013
Inv.	6,600	1,000	0	0
Personnel inc. SHI	1,188	7,450	8,150	8,200

Other running costs	1,132	820	830	840
Indirect	1,020	3,639	3,951	3,978
Cl.-ERIC contrib.	0	1,040	1,040	1,040
Total	9,940	13,949	13,971	14,058

Masaryk's University, Brno (Laboratory of natural language, Faculty of Information Technologies)

Masaryk's University Brno	Year			
Type	2010	2011	2012	2013
Inv.	1,000	300	0	0
Personnel inc. SHI	158	1,005	1,160	1,160
Other running costs	304	205	230	230
Indirect	185	484	556	556
Cl.-ERIC contrib.	0	0	0	0
Total	1,647	1,994	1,946	1,946

University of West Bohemia (The Department of Cybernetics, Faculty of Applied Sciences)

Department of Applied Cybernetics, UWB	Year			
Type	2010	2011	2012	2013
Inv.	1,150	300	0	0
Personnel inc. SHI	158	1,005	1,160	1,160
Other running costs	214	205	230	230
Indirect	149	484	556	556
Cl.-ERIC contrib.	0	0	0	0
Total	1,671	1,994	1,946	1,946

The Institute Of the Czech Language of the Czech Academy of Sciences

Institute Of the Czech Language, Academy of Sciences of the Czech Republic	Year			
Type	2010	2011	2012	2013
Inv.	1,100	300	0	0
Personnel inc. SHI	158	1,005	1,160	1,160
Other running costs	224	205	230	230
Indirect	153	484	556	556
Cl.-ERIC contrib.	0	0	0	0
Total	1,635	1,994	1,946	1,946

Investment costs are scheduled only for the first year of the construction phase (2011) at the total amount of 13.8 million Czech crowns for the network distributed infrastructure, servers and data storages of the LINDAT-Clarin Centre at the respective workplaces and the equipment for recording audio signal and electrical and electromagnetic signals for collecting data (language experiments and collecting data from human subjects). In addition to it, in the operational phase making use of the scheduled capacity for storing big language data by means of the CERIT project (requirements for storing hundreds of terabytes and 400-2000 CPUs (occasionally) for data processing) is expected.

- ***Proposed legal form***

At the first project phase (first three years starting in 2011, if it is not otherwise decided during this phase) the LINDAT-Clarin Centre will be a part of the Charles University in Prague which will arrange the cooperation of further subjects (Clarin, T4ME Net in the EU, the Masaryk's University, the University of West Bohemia and the Institute Of the Czech Language of the Academy of Sciences in the Czech Republic) upon bilateral contracts. There will be established separated workplaces (under the authority of the institutions, schools or faculties) with separated accounting which will implement the project in the respective institutions contributing to the project.

The Charles University in Prague will be as well entrusted with the enforcement of laws and responsibilities of the Czech Republic in the Clarin-ERIC consortium, if this organisation is set up in 2011.

After Clarin-ERIC coming into existence at the European level, the project partners in the Czech Republic will negotiate the suitable form of organising the node of the Czech Clarin-ERIC network; if no special organisational structure nor unit is created (e.g. in the form of interest group of legal persons), the Charles University in Prague will continue fulfilling this task. The national part of the project will be in any case executed by the respective workplaces of the partnership organisations, which will be true also after possible setting up Clarin-ERIC.

In case Clarin-ERIC is not established and Clarin coordination is arranged in a different way, the project partners are willing to negotiate about a different form of cooperation with the Clarin coordinating structure.

- ***Organisational structure and human resources in the project***

9.1. Organisational structure

The Clarin project suggests founding an association in the European legal framework of ERIC. ERIC, which is presently in the form of the EU directive (from 2009), expects establishing infrastructure associations, the members of which can be only national states or intergovernmental organisations or organisations delegated by these bodies. If the Ministry of Education, Youth and Sports of the Czech Republic transfers its executive powers in the future Clarin-ERIC to the proposed LINDAT-Clarin national Centre, the applicant is ready to fulfil these tasks.

A part of the project for the distribution of data is intended for the Clarin-ERIC contribution amounting approximately 10% of the national budget for this part. The amount will be specified in detail after founding Clarin-ERIC in 2010; the head office is going to be in the Netherlands (in the country where the coordinator of the ESFRI Clarin project has its seat in these days).

The tasks of the national project part will be carried out by the partnership organisations in accordance to bilateral contracts in individual organisational units with separated financing.

The LINDAT-Clarin Centre will set up the LINDAT-Clarin Centre Council with the responsibility to manage the activity of the centre and make sure that especially national part of the project (data collecting and annotating) corresponds to the current requirements of the existing research and development base in the Czech Republic in the field of linguistic and applied linguistic. The Council will have the status which will be approved at its constitutive meeting during 2010 (before starting the activity of the Centre). Council's resolutions will have the nature of recommendation for further operations of the Centre and will be presented as one of the factors of the Centre's evaluation by the provider. The Council will have 11 members, of which 7 members will be external ones (operating outside the LINDAT-Clarin Centre).

9.2. Human resources

Scientific coordination of the project:
Prof. RNDr. Jan Hajič, Dr. (the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague)

Administration, financial planning and managing:
Anna Kotěšovcová (the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague)

Professional guarantor of the project and coordinator of cooperation with Clarin and Clarin-ERIC:
Prof. Ph.D.r. Eva Hajičová, DrSc. (the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague Praha)

Scientific secretary of the project:
Mgr. Pavel Straňák

Responsible investigators of the project at the cooperating workplaces:

University of West Bohemia (Pilsen):
Prof. Ing. Josef Psutka, CSc. (Department of Cybernetic, FAS, University of West Bohemia, Pilsen)

Masaryk's University (Brno):
Doc. Ph.D.r. Karel Pala, CSc. (Faculty of Information Science, Masaryk's University Brno),

The Institute Of the Czech Language of the Czech Academy of Sciences (Prague):
Doc. RNDr. Karel Oliva, Ph.D.

Technical administration of the project, Clarin node (management):
RNDr. Milan Fučík (Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague)

- ***Legal form of intellectual property rights***

10.1. Open access

Due to the character of the humanities and computational linguistics which should be completely covered by the LINDAT-Clarin Centre in the field of language data, we expect providing the actual data and initial training from the funds of the LINDAT-Clarin Centre only for overhead costs. Remote access will be offered to registered users free of charge. Providing data for commercial sector will be governed by market conditions and a necessity to ensure an equal access to data produced from public funds within EU. All interested parties together with "data customers" will become members of the LINDAT-Clarin network or (and as the case may be) a successor of the Clarin project or a similar EU project in the future). The level of interest of users (a number of network users, a number of used data files, a volume of used services both locally and through "remote" services) will be considered by us a criterion for evaluation the activity of the centre (cf. Chapter □).

10.2. Licensing agreements

Licensing agreements are necessary to be divided into two categories:

- passive ones (collecting data to be distributed by the LINDAT-Clarin and Clarin networks);
- active ones (open licences to data produced in the national part, or possibly under the authorization of the Clarin network also for data produced outside the Czech Republic).

It is one of the crucial aims of both Clarin project and T4ME Net to make the data available to the broadest range of the users in an easiest manner not only in the technical sense. Licences for using the data in the non-commercial sector will be therefore stipulated so that the aim could be fulfilled, i. e. they will be open, enable further data modification under previously set conditions and support further free spreading such data by technological means analogical to the Clarin network. Such licence for the integration of data coming into existence both in the proposed LINDAT-Clarin Centre and externally to it, which would not disturb the above free access, will be required. Commercial licences will be awarded upon the equal treatment on the conditions, which do not eliminate small nor middle-sized businesses.

11. Human resources development

All participating institutions are eligible to perform educational activities at the level of Ph.D. studies; three of the partnership universities even have specialized programmes of master's type. The partnership organisation will also contribute to further specialized training by holding summer schools for professional public and especially by engaging students and professionals from practice to the activities of LINDAT-Clarin Centre:

- Theses and Ph.D.s;
- annotation projects (also for practical applications)
- exchange scholarships (LDC, ELRA/ELDA, …)

In addition to it, the centre will provide the training of the users in working with language data in a broad range, from direct initial training of users who otherwise require only remote access, through technological courses oriented on data application and basic software for the development of language technologies, to the inclusion of relevant themes to master's and

Ph.D. studies at the Charles University. The Centre will provide expert's opinions for both Czech and foreign students, scientists and researchers for their scientific and research projects. The Centre is supposed to be involved in the applied projects upon the direct contracts or through the applied research and development.

- ***Indicators for monitoring the project implementation – initial state***

The Centre of a type submitted by this proposal has not been so far operated in the Czech Republic or Europe (in any of its segments). The distribution is performed by standard methods (DVD, online one-time transfer, open source storage).

Language data and language software developed at all relevant workplaces in the Czech Republic have currently approximately 3,000 users at the total, about 200 users outside the Czech Republic (of them in the Czech Republic and abroad several software companies and language technologies companies such as Microsoft, Xerox RCE, IBM, Nuance, Aspi (Kluwer), Zica Corp. (Nokia), Umbria (Prentice Hall) a others).

A moderate growth is forecasted in the future, due to the extending the offer of data sources (besides other things related to the interconnecting within Clarin i T4ME Net) and their new forms of annotation (related to implementing in other branches – neurosciences, artificial intelligence etc.). In the view of the composition the biggest growth will be among the university students (including students from abroad) in the follow-up master's and Ph.D. studies; a higher interest is expected from smaller research teams of regional universities, humanities research workplaces and development teams from industry depending on penetrating "knowledge technologies" to the science and commercial world. The scheduled distributional capacity will serve as many as 5,000 users (several tens to hundreds concurrently for the remote access), and in the field of language data annotation approximately 6-8 concurrent projects based on the assignment or in relation to foreign research projects.

We propose to apply the following criteria for evaluating the LINDAT-Clarin Centre in the construction phase; if it is not otherwise mentioned, they relate to the sum from all of the cooperating institutions. If the initial state is non-zero, it refers to language data which have been presently prepared at any of the workplaces of the future LINDAT-Clarin Centre or data which have been already completed and are being distributed by one of the present publishers of these language data or directly any of the cooperating organisations.

Only data and data types that are going to be administered in the LINDAT-Clarin Centre irrespective of being created in the Centre with its financial funds or provided to the Centre from the projects presently running at the workplaces of cooperating organisations within other programs (Grant Agency of the Czech Republic, LC of the Ministry of Education, Youth and Sports, EU´ 6^{th} a 7^{th} Frawework Programmes etc.).

Similar criteria are to be applied for evaluating the operational phase (starting in 2014). Indicators values will be increased mainly in the area of data distribution because in the operational phase ("under standard operation") a significantly higher level of the LINDAT-Clarin Centre's utilization both by the Czech Republic and from abroad is considered. The values of the data production will be similar to those for the construction phase given in the table.

Category	Indicator, unit	Starting state	Annual volume / accrual during the construction phase of (average)	Planned at the end of the construction phase 13 (start of operational phase)
Human resources	Staff, employees with permanent contracts, FTE	0	N.A.	6-8
	Annotation teams, number of annotation groups	0	2	8
	Annotation manager, lingv._+ techn., FTE	0	2-3	12
	Annotators: running number (partly FTE)	0	10	40
Education and training	Number of students involved in LINDAT-Clarin, running number	0*	3	15
	Number of LINDAT-Clarin master's and Ph.D. studies graduates, annually	0*[1)]	0	2
	Number of seminars, schools and trainings for students, scientists and users, annually	0	N.A.	3
International cooperation	Number of international projects with direct participation of LINDAT-Clarin	2	1	6
	Number of study trips abroad (Clarin, other), annually	0	N.A.	10
	Publication activity (magazines, anthologies, books dealing with "language data"), number of publications annually	20*[2)]	0*[3)]	20
Data annotation	Number of collections in the database, public access, Clarin node, running number	0	N.A.	50*[4)]
	Number of starting units (words, state after collection and unification, all	0.5 milliard	100 million	1 milliard

	languages)			
	Number of annotated units (words, all annotation means, Czech language)	1.8 million	150 thousand	2.5 million
	Number of annotated units (words, all annotation means, other languages)	0.4 million	80 thousand	0.8 million
Data availability	Number of licences awarded for data administered at the Clarin node, total	0	5*[5)]	20
	Number of remote accesses to the data and node services, annually	0	0*[6)]	1000

Tables of indicators for monitoring the project implementation of the LINDAT-Clarin Centre within the construction phase

Notes:

*1) Number of students participating in master's and Ph.D. studies at four workplaces participating in the future centre is currently about 30, not all of them will, however, work on collecting, annotating and distributing data in the LINDAT-Clarin Centre. The number of graduates ranges between 6 and 8 annually, and only some of them from the existing workplaces will join directly the LINDAT-Clarin Centre. We rely mainly on engaging newly enrolled students of the regular studies.

*2) Publications are included in ISI Thomson lists and in the list of outstanding non-impacted publications dealing with the themes that will be solved in LINDAT-Clarin in the future.

*3) A significant increase of the above specified type of the publications is not expected in the Centre itself. However, an outstanding increase of publications that originated from the processing of data produced in LINDAT-Clarin is anticipated. Unfortunately, this number will be hardly identifiable because the citations are written by the authors outside the participating workplaces; we suppose this number will be around 50 publications a year.

*4) It should be noted that the data within the construction phase will be provided only during the testing operation; their number will increase substantially after 2014 (full operation / operation phase).

*5) The number of awarded licences for the data will be unlimited at the construction phase.

*6) Number of accesses is expected to grow to the scheduled capacity (tens of thousands of accesses) at the full operation in 2014 and later as well.

- ***Socio-economic impact analysis***

The LINDAT-Clarin Centre, though it is designed as a servicing research infrastructure, will have several important socio-economic functions even outside science and research fields. These are as follows:

- It will provide education in the perspective scientific discipline, which will increasingly assert itself in the "knowledge society", providing education not only by lecturing method, but also through direct working at annotating and software projects that will be applicable in practice straight away;
- it will reduce so called entrance barrier for small scientific teams which cannot otherwise afford to organise collecting and annotating language data or pay high prices for their even non-commercial licences relating both foreign and above all Czech data; it will as a result enable to develop language software applications which are otherwise necessary to be collected abroad and are not always at the best quality;
- it will enable the Czech Republic to participate in the EU projects (both at basic and applied research), where the existence of language data is expected or explicitly required (e.g. projects for machine translations) and without them the Czech institutions would not be accepted in the investigators´ consortia, which proved to be true in the past;
- it will enable to preserve cultural language heritage (as a part of the project for making available the archives resources of the Institute Of the Czech Language of the Academy of Sciences of the Czech Republic);
- in the European context, "make visible" the co-working institutions and allow them to apply also for other EU projects both in basic and applied research;
- it will enable as a result maintaining Czech language as a national language with the aim to raise it to one of frequently used communication means in the CEE region.

We do not specify in the above list direct impacts on the employment rate of the graduates of the follow-up master's and Ph.D. studies, which are even more important because the work character of language data collecting and annotating gives work to some categories of employees which are generally considered disadvantaged; it relates to part-time work or possibly "remote work" (home office) (relating to parents that look after small children who are relatively numerous in this branch, handicapped people etc. – such solutions have been used in the institutions applying for this project for quite a long time).

- ***Other facts***

The Charles University in Prague as an applicant is a public university and as such it does not need to justify that it is a research institution according to the article of 30 Sec. 1 of the Commission Regulation (ES) 800/2008.