Integrated Strategic Plan for Supporting HSS Research

DELIBERABLE 3C-6.1

16-05-2011

Editors: Tamás Váradi, Piroska Lendvai

www.clarin.eu
The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.
Integrated Strategic Plan for Supporting HSS Research

Deliverable 3C-6.1

CLARIN-2011-05

EC FP7 project no. 212230

Deliverable: 3C-6.1
Extended deadline: 30.04.2010

Responsible: Tamás Váradi

Contributing Partners: HASRIL, Göteborg University, Bergen University, Institute for Psychological Research of the Hungarian Academy of Sciences, Université de Strasbourg, Université de Nancy 2, Huygens Institute, Lancaster University

© All rights reserved by HASRIL on behalf of CLARIN
Scope of the Document

The current deliverable provides an overview of the results achieved by the activities implemented during the period of a one-year collaboration within CLARIN-supported projects, between CLARIN experts as advisors and Humanities and Social Sciences researchers as users. One objective of the collaboration was to help researchers manage and enrich their own digital research data and thus to advance scholarship by means of enhanced, automatized curation, analysis, or interpretation of digital texts. To this end, CLARIN provided knowledge about and expertise in applying LRT tools for linguistic and semantic processing of data sets from new domains. Another objective was to raise researchers’ awareness in linguistically enabled research and to impact the target group via real-life use cases. The strategic suggestions put forward in this document aim to guide providers of linguistic resources and infrastructure services in interacting with their target audience in collaboration scenarios.
Table of Contents

1. Introduction ...............................................................................................................................................6

2. Bringing LRT infrastructures closer to HSS scholars ................................................................. 7
   Communication across communities ........................................................................................................ 7
   Cross-disciplinarity ............................................................................................................................... 8
   Transfer of knowledge .......................................................................................................................... 8
   Further, continuous training and education ..................................................................................... 10

3. Implementation .....................................................................................................................................11
   Workflow ................................................................................................................................................ 11
   Step change in mentality ....................................................................................................................... 12
   Joint development of LRT ..................................................................................................................... 12
   Consolidating the impact .................................................................................................................... 13
   Suggestions for CLARIN .................................................................................................................... 13

4. Contributing to the Humanities BLARK ..................................................................................... 14
   Awareness in the Humanities BLARK ............................................................................................. 14
   LRT utilized in the collaboration use cases ........................................................................................ 15
   Validation of LRT applied in the use cases ....................................................................................... 16

5. Publications ............................................................................................................................................19
   Joint publications of researchers and advisors .................................................................................. 19
   Researchers’ publications ................................................................................................................... 19
1. Introduction

One of the objectives of the CLARIN project’s preparatory phase was to build bridges to its target user group: the Humanities and Social Sciences (HSS) research community. To realize this, the period of a one-year collaboration was offered by CLARIN to HSS scholars in order to help further articulate the relationship between new technologies and humanities. CLARIN delegated experts in Language Resources and Tools (LRT) as advisors to teams of HSS scholars, and provided practice for the researchers in understanding the nature and means of using LRT, desirable to adopt for the particular objectives of a handful of ongoing HSS projects. Each supported project aimed to benefit scholarly research in a knowledge institute by means of incorporating LRT in preparing data or tools to be used for scientific investigation. The following HSS disciplines were addressed in the projects:

- Linguistics (domain-specific language use; dialectology; lexicography; lexical semantics)
- History
- Cross-disciplinary research involving History and Psychology (on historical narratives)
- Cross-disciplinary research involving History and Philosophy (on the circulation of knowledge).

In turn, CLARIN has gained first hand knowledge about the research interests, methods, and conceptual frameworks based on which HSS scholars conduct their research. This knowledge is the key to tailoring the services of the CLARIN infrastructure to suit the needs of its target communities.

Besides the community of humanists, the targeted user group included the IT personnel of HSS researchers, as experts possessing the skills for developing and delivering digitized data, on which the scholars would want to perform automated analysis and enhancement. Profiles of the users necessarily varied; typically, the research teams had possessed very limited prior knowledge of LRT before the start of the cooperation: they were able to formulate valid needs with respect to the required output, but — even the IT staff— could be easily mistaken concerning the necessary amount of work, the kind of tools and workflow, and the actual implementation or methodologies the project needs would imply.

In the project use cases, practical, often complex goals were defined: storing, accessing, and/or deriving information from humanities data. Details of the collaboration projects are thoroughly described in CLARIN Deliverable D3C-5.1. On top of the continuous monitoring, after the collaboration period, the CLARIN WP3 has gained additional structured feedback from the participants by means of a questionnaire survey, so that the course, experience, and output of the collaboration could be assessed, and documented. Furthermore, opinions, both critical remarks as well as positive impressions were collected, so that important issues, factors, and facets of the interaction between the two communities could be identified. Most of the questions in the questionnaire solicited information in free text form from the respondents. Ten persons have completed the form: six researchers, and four advisors.

The recommendations laid out in the current report are distilled from personal experience, close interaction, the questionnaire, as well as previous documents of related CLARIN WPs. They pertain to general issues around the transfer of LRT expertise (i.e. knowledge input and output), as well as around important facets of the interaction between the two communities (such as the course, experience, and impact of the collaboration). Based on the experience CLARIN has gained, the deliverable reports on important facets of two major issues:

- Bringing LRT infrastructures closer to HSS scholars, and
- Findings concerning implementation and related technicalities.
2. Bringing LRT infrastructures closer to HSS scholars

The experience gained during the collaboration period concludes and emphasizes an important, straightforward, but — in a computational setting — non-trivial finding that the best impact on non-technical scholars can be achieved by hands-on exercise. This outcome is supported by previous external studies; for a survey see Sections 2.2 and 5.3 of CLARIN D3C-3.2. The promotion of the CLARIN research infrastructure could be most effectively carried out by the joint work of the communities of LRT experts and HSS researchers. In this way, research teams were acquainted with and applied methodologies of LRT in the realm of a specific HSS research project, learned about basic best practices, located and utilized basic computational resources for text-based research, and, in improved cases, adapted and extended language- or domain-specific tools as required by specific needs.

- Personal exposure to LRT ensures the elimination of a significant initial barrier, that of getting into the realm of a different discipline.
- In the CLARIN collaborations, the guidance provided by advisors promoted the acceptance of LRT to the HSS scholars and their IT team (if available) as a supportive and innovative research approach.
- It made the researchers realize that LRT is an agglomerate of acquired skills, methods, and ways of thinking.
- Based on the collaboration, users were able to identify basic linguistic core functionalities that constitute the ground for performing the more complex tasks required by HSS research.
- Importantly, the research projects are to extend beyond the time of the CLARIN advisory period, which allows for more establishment and dissemination of the accumulated know-how.

Communication across communities

- Feedback received from both communities clearly stated that for effective communication, allocating a large amount of time is required (which presupposes the availability of financial resources). Meeting this condition enables the guidance of novice users in accessing LRT technologies and resources.
- Regularity of contact between advisors and researchers is important: sporadic information exchange does not suffice.

In the verbally active CLARIN use case projects communication via email, phone and personal meetings often took place on a weekly basis, or even several times a week in the most dynamic weeks of the project. Specially focused tutoring/guidance/workshop sessions were however rarely organized, except for one project where three internal workshops could be arranged, gathering all the research team, the group of advisors, and even some external experts. Whether this is a possibility, depends on geographical distance and financial resources, among others.

When ranking the usefulness of different modalities in communicating, different views can emerge based on the project needs, in particular:
- Workshops are most effective
- Phone calls are necessary sometimes when discussing decisions, and at the start of the project (in order to explain what the research is about)
• Face-to-face communication is prioritized over email
• Face-to-face communication is an important and valuable as point of departure. Email-communication is sufficient in later stages – as a general rule
• Face-to-face meetings are useful but it may be difficult to find the time or justify the costs
• E-mails turn out to be useful for the following reasons: both parties are likely to have a busy work schedule, and e-mail communication is less interruptive than telephone calls; written explanations and elaborations can be studied at leisure, and serve as a documentation, which is not the case for oral input. Email is also motivating because in writing one has to organize ideas and present them in a structured manner
• A face-to-face meeting is needed to discuss complex items such as a demonstrator.

Cross-disciplinarity

Cross-disciplinarity may progress only slowly. Joint work entails the influx of new concepts and approaches to each other’s everyday work. Informing each other is a long and never-ending process. Experience shows that it can be hard enough to learn to work with a dozen new (complex) concepts from another field. Crossing the borders need to be done with care, in a limited, and in any case non-intrusive manner. It can happen that even after a whole year of collaboration, HSS specialists will use a natural language processing (NLP) term erroneously, or that they still cannot comprehend the use or the limitations of one or another technique — or vice versa, NLP advisors can overlook important properties of an issue to be formalized.

• It is important to acknowledge that LRT has not come into being as a supporting discipline of Digital Humanities (DH), but it is applied NLP, a complex discipline that blends several fields of engineering and (corpus) linguistics. It has evolved independently, and is therefore not intuitive to learn and use.
• Utilizing LRT, as opposed to using infrastructures that exploit LRT, may imply different processes and goals.
• It is to be established whether a project aims at developing tools (single applications) and pipelines (integrating/reassembling applications), or it can effectively make use of existing research infrastructures (web services that combine applications or workflows, typically tailored to specific research).

In order to understand what is DH, and what counts already as NLP, several activities can be of help:
• Learning about relevant previous projects and tools
• Familiarizing with digital repositories and the LRT they contain
• Learning to read documentation
• Learning to generalize from previous work, even if in a different field
• Developing a feeling for portability of approaches.

Transfer of knowledge

The digital maturity of the HSS researchers and IT teams was characterized by CLARIN advisors as low in some important areas, which was certainly anticipated from previous experiences. One negative remark pointed out that the HSS people in the team used obsolete technologies and ‘antique’ computer platforms.
The advisors:
- experienced that the amount of time that was invested in counseling was perhaps too short (due to a general shortage of time to spend on the project),
- in other cases they felt it was taking a bit long, but was necessary, since HSS researchers often need time for technical issues to mature, and for HSS data to be brought into a proper form;
- in one case the process was judged too long, due to unforeseen personnel changes.

Concerning the collaboration, it was investigated whether LRT expertise and knowledge was transferred to HSS and IT specialists in general, for which positive evidence was found. Compared to the initial situation, researchers developed significant knowledge about LRT, so that they were increasingly able to think in term of LRT categories. A major expected impact of the collaboration between HSS projects and CLARIN was that the specific LRT expertise to which the HSS researchers are exposed shall bring novel insights both with respect to the actual research data, but also data analysis methodology. Indeed, researchers
- reported on progress in their own discipline
- enthusiastically incorporated new methods and technicalities in their everyday research
- claimed that the collaboration acted as a catalyst of major individual achievements and professional growth
- exemplified the impact of transdisciplinary approaches that can yield progress and catalyze a paradigm change in HSS research.

Services that form the core objectives of CLARIN can be provided optimally when LRT needs of HSS specialists are thoroughly considered. This enables advisors to better guide experimenting and provide for innovation in HSS research.
- The original HSS projects’ descriptions of work were often prepared together with the advisors, so that there was a quite good common understanding of the goals of the project, or they were fairly clearly written by themselves. Nevertheless, long discussions on needs and possibilities were necessary to clarify things and to divert activities to follow a more practical line. Often, the focus needed to be shifted to structuring, formatting and encoding issues, instead of analysis.
- Advisors feel to have provided a limited introduction to the LRT field for the project team, so that the researchers could gain insight into certain issues (e.g. structuring and xml, statistical methods).
- As expected, the advisors learned about research methods and types of resources utilized in typical HSS projects, and developed more understanding of the LRT needs and capabilities of the HSS community. From insight into fieldwork, they gained added benefits: there was an increase in the understanding of HSS research issues (one advisor added: “even though I have worked with computing in the humanities for almost 35 years”).
- They were however doubtful if the outreach activities had a satisfying level of diversity and did not feel that counseling could be propagated to the project team’s institutional level.

In investigating room for improvement, unforeseen bottlenecks may be encountered during the collaboration, in particular:
- Lack of time
- Lack of structured collaboration
- Lack of adequate communication
Personnel change
Initially different notions of what an application should achieve or what kind of functions are needed.
These can be alleviated by more communication and redefinition of some of the goals.

On assessing whether the collaboration yielded cross-community engagement, respondents report that

- They will apply for funding based on the results of the project
- The project federated several people working in Linguistics, Translation and Corpus Linguistics. It opens a new research theme, which in the future might be a kernel for creating a new research team.
- It has resulted in one joint publication at a LRT conference, but at least one more is planned in a HSS forum.
- Researchers become more aware of the possibilities offered by newer methods and tools.
- New research lines and involvement of students may emerge that are fully based on the use of language technology and corpus linguistics in a specific HSS discipline.

Impact can hereby propagate beyond a single project, for example to the project’s host institution. Respondents experienced impacts such as:

- “I think it has made a big difference in the way my research is now perceived because of the European scope of this cooperation, and also because of the use of LRT in my research. I learned new skills in a very specific manner, based on my research.”
- “We are realizing that utilizing this kind of primary data in this way has a good deal of ‘novelty value’, i.e., is considered methodologically leading-edge among other historians.”
- “More knowledge of and expertise with LRT yielded the emergence of similar projects, building further on the results.”

Further, continuous training and education

Awareness that has been raised in LRT matters might entail that HSS scholars express their need for further education, which could accommodate different modalities of training and education. In some cases users can have less, in others higher expectations than LRT can provide for. The complexity of this issue however requires more structured investigation to understand motivation, false presupposition, and fight the frustration of users who might experience some disappointment due to unrealistically high expectations about the current potential of LRT and integration of information science and information technology in the humanities. Importantly, they become conscious that linguistic analysis is a prerequisite of higher-level content processing.

According to the survey, users do not regard the time that was necessary to select and create tailor-made tools as ‘wasted’, but rather as an investment that adds to the general knowledge of each field. The researchers experienced the amount of time invested in learning as normal, but the most advanced of them acknowledge that it can be part of a very long, continuous and unfinished process. Some of them sense that it was quite long to become aware of existing resources. They claim to have obtained a basic introduction to the LRT field, which makes them feel learning more, realizing its potential. Collaboration ensures personal success of both early- and advanced-stage researchers. In accordance with the planned Knowledge Sharing Infrastructure CLARIN has been aiming at, individual coaching could be enhanced by the CLARIN infrastructure including services such s
Question Answering and Helpdesk systems, as well as the web-based Community Creation and Consolidation Service of CLARIN (see CLARIN WP6 deliverables).

3. Implementation

As a major impact of applying LRT infrastructure in HSS disciplines, users will experience the ‘evident’ added value in their daily research such as improvement over manual content analysis and generation: speed, efficiency, and the ability to replicate output. They might grow to become engaged in digital humanities, the emerging discipline, and increasingly make use of new ways of research and collaboration, as well as cross-fertilization between their original discipline and LRT. Members of the research teams may be interested in having a role in development of content and goals for LRT, both in the advisory and end user capacity.

Workflow

As a result of the collaboration, workflows in the use cases tended to shape up in meaningful ways.

- Researchers gained awareness about setting up preprocessing steps, as well as stopping criteria, in their work. To apply statistical methods with success and reliability, significant volumes of data are needed as input to computational systems, and they need to be prepared first. Preprocessing turns out to be extremely time-consuming due to specific types and sources of data in the humanities.
- Technical issues such as data preparation and processing can be learnt to be consciously prioritized over the interpretation of the analysis results. The ranking of the priorities within the project goals (i.e. which tasks were treated as most/least important) as specified by the researchers themselves was varied. Typically, the highest priority was assigned to
  - developing new methods for exploring texts for specific research (data interpretation and visualization tools, linguistic description),
  - building domain-specific corpora,
  - pre-processing (spelling normalization, stemming, lemmatization),
  - metadata acquisition,
  - developing specific lexical categories in a specific language,
  - developing lexical and syntactical search algorithms (graphs),
  - clustering approaches for content analysis in a specific software.
- In other cases, projects from the very start targeted pure technical development, e.g. replacing outdated pieces of software. In these cases, priorities included receiving advice in the process of finding new software systems that fulfill the special needs of the project (for instance special keyboards, the creation of new fonts, etc.), and analysis of language data (format, structure) to address solutions regarding transfer to new programs.
- Least attention was reported to having been directed at doing fundamental HSS research, or improved specific technical developments, such as the definition of a unified query language.

During the collaboration, experts were often invited to participate in evaluation/internal assessment within certain phases of the project (typically: midterm and end reviews). However, it occurs that occasional revisions in the course of the project work were more typically done not based on the reports, but on the feedback given at each interaction, on a more continuous basis. For example,
when consultancy was provided on tools, it gave more emphasis to language technology, and has altered the workflow. Some of the revisions can take place at a quite early stage and thus change the order in which a researcher works on the project goals. For example, it may come as a revelation that even for running simple statistical tests on manually extracted variables, the full dataset involved in the research needs to be formalized (e.g. by XML encoding), including manual checking and correction, whereas the researcher presumed that this would be necessary only if specific corpus analysis would be done (such as concordancing).

Step change in mentality

- Ambitious users need to be warned about the limits of addressing complex or abstract HSS research issues, because these pose serious challenges to LRT infrastructures, being difficult to operationalize in LRT terms.
- On the other hand, providers must learn to cope with a changing view on infrastructure building. User needs ask for maximizing applied utility, allowing flexibility for tailor-made solutions, and preparing accessible documentation.
- Providers should not expect users to develop the ability to think in terms and methodology of LRT, and turn into experts (otherwise they would already qualify as trained NLP specialists).

Joint development of LRT

- For optimal collaboration that involves cross-fertilization among the communities, a balance needs to be reached between the scholars’ intuitive methods and the results yielded by algorithmic approaches in LRT.
- It is recommendable that users are made aware of the issue of formalization: formalizing a task means creating structure and turning characteristics and values that are hidden in texts into explicit properties.
- Users need to realize the importance of structured representation of meaning: to create means for structured resources for their own research area (e.g., glossaries, thesauri, databases)

Even though the HSS teams prefer applying simple tools to their data, they are likely to come to an understanding of the utility of more sophisticated tools (although from their point of view such tools, too, should remain ‘simple’ – in the sense easy to use and intuitive). Their opinion is that tools should always be simple even if what they achieve requires sophistication.

In CLARIN, the stimulating and interesting points experienced in joint development have turned out to be

- Learning from linguists about corpus linguistics, language technology and ontologies – this knowledge is now gradually becoming part of the involved HSS field
- Having an external point of view to the HSS project; being aware of some new LRT available from CLARIN
- Learning more about LRT and statistical tools with specialists in a tailored manner. It also forces one to be more precise about one’s own needs for the HSS research, generating added value to the presentation of research results
- Specific input to specific needs of a project, clearly stated and based on concrete analysis of tools.
The most inspiring experiences that catalyzed processes were listed as
- Face-to-face meetings/question-answer sessions
- When the researcher got new insights
- The simplified encoding
- The decision of building one's own corpora after an evaluation step
- The first workshop, which resulted in a useful view on the development track
- The reliability of the data extraction
- The fact that the methodology could be used again, and be improved.

**Consolidating the impact**

Good means of documenting the process and expertise gained from the partnership are
- The writing of project reports,
- Documenting demonstrator applications, as well as
- Collaborating on articles.

For maximal impact, it is highly important to disseminate the output more widely, and allow the replicability of the work that has been carried out by means of describing its technical details. In some cases, the increase of staff expertise was also (in the process of getting) documented, e.g. by reports and manuals.

**Suggestions for CLARIN**

Respondents of the questionnaire survey were asked to provide suggestions with respect to the collaboration and outreach activities of CLARIN. The comments were the following:
- Advertise more the available common resources
- Provide training (for XML for example)
- Help publicize
- Provide access and room for centralized servers
- Focus on sufficient time and communication between partners
- Simplify the access to existing resources (for example in order to utilize the Weblicht platform: Several colleagues from HSS will be interested in using these tools for their own projects. At the moment, it is necessary to get a password to access the platform)
- Allow space for small experiments
- Funds from CLARIN to support HSS projects (as end users of the resources and tools)
- Of course more money would allow more and better collaborations
- Perhaps offer basic skills for project staff, e.g courses in XML and statistics.
4. Contributing to the Humanities BLARK

To avoid the development of resources for each separate Humanities discipline, the so-called Humanities “BLARK” (Basic LAnguage Resource Kit) is initiated and compiled by CLARIN WP5. It indicates which resources are necessary to do research in some specific SSH field. The Humanities BLARK defines, links, and documents HSS user requirements with respect to specific LRT applications, and identifies the modules and data that are needed to build those applications; cf. Section 3 and Appendix A of CLARIN D5C-4. In the CLARIN collaboration projects, feedback on issues pertaining to the Humanities BLARK has been collected from the HSS users via the questionnaire survey. The results are set out below in the form of replies to six directed questions (taken from the WP5 deliverable).

Awareness in the Humanities BLARK

1. What are the key language resources/services that your field needs, and how would you use them in your research?
   - Digitized older texts
   - Language tools for all stages of all languages occurring in my data set
   - Word frequency lists, automatic creation of graphs and trees
   - Corpus managing tool, Dictionary editing system: XML-editor; XML-database; flexible installation, set-up and maintenance in order to meet the specific needs of the project; standard formats for flexible and reliable import/export
   - We need both raw and annotated corpora. Annotation might concern: tagging, lemmatization and syntactically annotated corpora. Even semantic annotation will be a good thing (frames from FrameNet). Web services such as those provided by Weblicht are very useful: tagging, lemmatization, parsing, named entity recognizer, available for several languages
   - Corpora (raw, tagged), concordancers (on-line, with a simple query language/interface), tagging and lemmatization tools. Ontologies and lexical database (FrameNet or WordNet)
   - Spelling normalization to enhance analysis by better preprocessing results
   - NER: extracting social and geographical information from texts

2. What are your text-related research issues and tasks that are instrumental in achieving (research) goals?
   - Text search and browsing, text selection
   - Encoding text, selecting diagnostic labels (terms)
   - Managing of dictionary data
   - We use corpora to check our linguistic hypothesis, to be able to give a complete linguistic description of a specific phenomena
   - Retrieval of linguistic information in a text corpus as a tool in the editing of the dictionary

3. Which of these tasks could be (partly) automated?
   - Probably all
   - Certain parts of the editing process (for instance control of references)
   - Importing citations directly from corpus managing tool into dictionary edition system
   - Collecting data from corpus should be optimized. Instead of learning a corpus query language, it would be nice to formulate a unique query and this query would be sent to all the
existing corpora available from the CLARIN repository

4. What besides text is important for your daily research life?
   • Historical artefacts and probate inventories
   • Research literature, Google, domain-specific forums
   • Creating diagrams and specific trees and charts to prove my point (onto-terminology?)
   • Editing is based on a collection of paper slips, and the corpus is used as supplementary data
   • Metadata

5. What would you expect a computer to do for you sometimes?
   • Our activity is fully computer-based, from document retrieval via data generation to statistical data processing
   • Take care of trivial work routines
   • Assist in retrieving and processing more or less complicated data
   • Be able to do explorative work in corpora
   • Do statistics (trivial and explorative)
   • Handle, organize and control data
   • Linguistic annotations

6. What tools in your field are you aware of and/or are you already using?
   • Concordance tools, KWIC lists, simple correspondence analysis using R
   • NooJ
   • Excel
   • SPSS
   • Concordancers, taggers, lemmatizers.

LRT utilized in the collaboration use cases

Data resources:
   • a digitized text archive (corpus) of 19th and early 20th century fiction
   • an associative thesaurus
   • a digitized dictionary of 19th century language
   • dictionary material
   • transcript of spoken material (corpus)
   • national corpora
   • metadata about the informant whose data is included in corpus development (place of residence, age, occupation, etc.), the interviewer and the subject of conversation
   • lexical databases
   • FrameNet
   • numerous multilingual/monolingual corpora

Technology that was employed as part of the applications:
   • POS tagging
   • lemmatization
   • named entity recognition
   • a dictionary editing system
Some of the tools used were freely available, the majority required adaptation, or was assembled from pieces used in other projects. Tailor-made programs and stylesheets were developed from scratch. The proportions are difficult to judge, the projects were utilizing a mix of solutions, which is attested below.

LRT created within the project, developed from scratch:
- (parallel) text corpora
- extraction patterns
- xml-diagrams using metadata
- a site was created to display the data
- a computational morphology module for 19th century Swedish
- automatically linked entries across dictionary entries
- word lists/gazetteers
- metadata
- search algorithms
- user interface

Modified/adapted existing resource:
- creation of gazetteers from bigger corpora
- keyword analysis
- topic modeling
- n-gram based language recognition
- visualization
- spelling normalization

Sometimes, combinations of applications were created for annotating the corpora (POS tagging, lemmatization, NER annotation, sentence alignment), and for preprocessing (tokenization, language recognition, stemming, topic analysis, visualization / user interface pipeline).

‘Non-standard’ design choices were also employed, e.g. to encode data in XML but not according to any standard markup schemes, since no suitable encoding could be found. (It allowed the researcher to do the work faster. Afterwards, the simplified encoding was automatically converted to XML with a tailor-made program, with a little bit of manual correction.)

**Validation of LRT applied in the use cases**

Scholarly activities that have been conducted can be characterized in the following dimensions:
Quantitative research carried out in terms of:
- Most frequent word and their contexts, their frequent collocates
- Statistical analyses run on the tables of metadata and term data
- Term occurrence counts

Qualitative research issues addressed:
- Finding out diagnostic labels (terms) used by speech and language therapists
- Extracting information from a corpus of transcribed tape recordings in a dictionary project. All kinds of linguistic information are of potential use in the entries of the dictionary: phonetic, morphological, syntactic, semantic, pragmatic, which is not normally attested in the collection of paper slips
- Topic analysis
- Search for named entities pairs (Person, Person) (Person, Organisation), (Organisation, Organisation)
- Thesaurus-expanded browsing of 19th century fiction in order to answer a historical research question
- Subjective time experience as a personality trait
- Subjective time experience as marker of depression
- Social representation of the events of Hungarian history
- National and European identity
- Historical trajectory research
- Distant monitoring of psychodynamics of isolated small groups
- Space psychology (content analysis of crew reports in space-analogue experiments)
- How knowledge was spreading in time and space

Processes that could additionally be automatized include the following:
- the generation of semantically similar search terms
- input of new documents into the system with XML encoding
- a single query language to search simultaneously all the existing corpora
- corpora should be selected by faceted search provided by CLARIN repository.

A demo/prototype system has been (partially) created during the projects in several cases, or will be created in a next stage. These feature the following functionalities:
- Solutions to the specific needs of the project, for instance the possibility of displaying full text and printing of paper slips with relevant occurrences of words, including basic information and a certain amount of surrounding text around the words
- Provide a list of patterns available in Unitex format (FR, DE, EN, RO)
- Various modes to select documents; facets: topic, person, location, language; visualizations: word cloud, time line, map, social network; provide links to transcription of the documents.

Input from the advisors that is necessary to implement, but requires future work:
- Web based concordance
- More work on subdocument segments
- Maybe also the creation of a common platform where new documents could be added
- Semantic Role Labeling
- Transfer of site and software to the researcher's institution
- Spelling normalization
• Further research on topic modeling, keyword and concept extraction.

General comments about the specific aspects of the collaboration:
• We are very glad to see that there are massive efforts in Europe to break the English language domination
• In the hope that our cooperation will continue for many more years, we are happy to devote our efforts and to contribute to this great endeavor
• I think I was very lucky to obtain such help in my research, I think I made the best out of it and worked hard to improve on my first idea. It helped me go further. However the time necessary to prepare my reports for the encoding could have been reduced had I known more about the available tools prior to the CLARIN cooperation
• We are in the fortunate position that this has been the beginning, rather than the whole, of our collaboration, so that we will in fact still be able to pick up on some of the experiences from the project in our planned future work.
5. Publications

- The collaboration resulted in publishing the results at national and international conferences.
- It also yielded professional contacts with new scholars interested in the methods and output of the projects.

Joint publications of researchers and advisors


Researchers' publications

• Vincze, O., László, J. (2010). The role of narrative perspective in history schoolbooks. Magyar Pszichológiai Szemle, 65, 4:571-595. [In Hungarian]