User-driven Language Technology Infrastructure – the Case of CLARIN-PL

Maciej Piasecki

G4.19 Research Group
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
maciej.piasecki@pwr.edu.pl
www.clarin-pl.eu

Abstract

The paper discusses a user-driven development of CLARIN-PL, the Polish branch of the European language technology infrastructure for Humanities and Social Sciences. CLARIN-PL can be used as an exemplar of a bi-directional (i.e. top-down and bottom-up) approach to developing language resources and tools. The paper presents an overview of the state of the basic processing chain for Polish, the set of basic Polish language resources and tools and typical processing schemes emerging from the development of key applications. We also discuss the problem of the quality of services offered by language tools that goes much beyond the typical measures used during testing. In conclusion, we try to envisage further user needs and further language technology infrastructure development for which the 3-4 year construction phase is a good starting point for a fully-fledged infrastructure.

Uporabniško usmerjena jezikovnotehnološka infrastruktura – primer CLARIN-PL

Prispevek predstavi uporabniško usmerjen razvoj CLARIN-PL, poljske veje Evropske jezikovnotehnološke infrastrukture za humanistiko in družboslovje. CLARIN-PL lahko uporabimo kot primer za dvosmeren (tj. od zgoraj navzdol in od spodaj navzgor) pristop k razvoju jezikovnih virov in orodij. Prispevek poda pregled stanja osnovnega zaporedja obdelav jezikovnih podatkov, množico osnovnih jezikovnih virov za poljski jezik in orodij ter tipične sheme za obdelavo, ki izvirajo iz razvoja ključnih aplikacij. Prispevek obravnava tudi problem kakovosti storitev jezikovnih orodij, ki presega tipične ukrepe, ki se uporabljajo med testiranjem. V zaključku je podan oris nadaljnjih potreb uporabnikov in razvoja jezikovnotehnološke infrastrukture, za katerega je 3-4 letno obdobje izgradnje dobra osnova za popolnoma izdelano infrastrukturo.

1. Introduction

Language technology infrastructure (LTI) is a complex system that enables combining language tools with language resources into processing chains (or pipelines) with the help of a software framework. The processing chains are next applied to language data sources in order to obtain results interesting from the perspective of research needs of different groups of users.

Addressing user needs is the basic challenge in software engineering. Users make all systems imperfect, but the truth is that software systems do not exist without users. They simply do not have a purpose. Moreover, LTI is interesting only when its proper users are significantly different from its constructors, as it would be good to finally see language technology (LT) going beyond the level of toy systems. Basically, language engineers should not construct LTI mainly for themselves.

Users should be present at all stages of system development. In a user-driven system development process, the Context of Use[1] determines the perspective from which the users perceive LTI. Usability (defined in terms of *efficiency*, *effectiveness* and *satisfaction* (ISO, 1997 1999)) is the basis for the assessment of any interactive system including LTI.

In this paper we will discuss consequences of the user-driven development for LTI construction. We will focus on the exemplar of CLARIN – a European LTI which is meant to support researchers in Humanities and Social Sciences (H&SS). CLARIN intended users are significantly different from its constructors and usually do not possess any knowledge of computational linguistics or programming.

2. Language Technology Infrastructure

LT has been developed for more than 10 years now. LT originated from the change of small limited systems characteristic for early NLP into robust text processing technology based on sets of exchangeable and reusable components: dynamic – language tools and static – language resources.

The idea of LTI comes from the observation that we can identify several barriers that prevent wide spread use of LT outside the world of computational linguists and computer scientists, cf (Wittenburg et al., 2010), namely :

- *physical* – language tools and resources are not accessible in the network,
- *informational* – descriptions are not available or there is no means for searching,
- *technological* – lack of commonly accepted standards for LT, lack of a common platform, varieties of technological solutions, insufficient users' computers,
- related to *knowledge* – the use of LT requires programming skills or knowledge from the area of natural language engineering,
- *legal* – licences for language resources and tools (LRTs) limit their applications.

LTI is a complex system providing a technological platform for the integration of different LT components into

[1] Context of Use encompasses users and their characteristics relevant to the general goals of the future system, users' tasks and their effects and different kinds of environment (technical, organisational, social and cultural).

one interoperable system. Moreover, other aspects like legal and informational ones are also taken into account.

CLARIN is a LTI focused on the use in the area of H&SS. The main goal of CLARIN is to decrease the barriers, as far as possible in the context in which LT is used by researchers from H&SS.

3. Development Schemes

CLARIN[2] is being built by an ERIC consortium of several countries that are obliged to contribute parts of the LTI. Different countries follow different schemes, however some common features can be identified. There are two possible basic schemes. The first is a *bottom-up process*, which can be also termed a *collected offer*. It is based on linking the already existing LTRs, and it is focused on establishing accessibility and technical interoperability of LTRs, as well as on establishing a common system of IPR licences that lowers the legal barrier. A distributed authorisation system is introduced and federated search mechanisms for searching the content of the resources and metadata in pre-defined formats (Wittenburg et al., 2010) are proposed. As a result, the tools and resources will become accessible via Web to the users and can be combined into processing chains. The only question left open is if the users know what to look for and what to use. LRTs mostly require from users the specific background knowledge, e.g. complex Slavic tagsets. LRTs often seem not to be directly related to the research performed by the users from H&SS.

Web applications both for individual services and for adaptable workflows for natural language processing for final users are mostly on borders of the main focus of CLARIN. *Usability aspects* (ISO, 1997 1999) and especially *usability evaluation* of the applications are very often neglected. At the same time, data presentation in resources and the results of processing in tools are implemented according to the user needs that are unknown! Processing chains are adapted to the unknown user tasks, whose goals mostly go beyond the domain of natural language engineering. However, at the same time, LTI is a new enabling technology that can create new needs, if well presented and explained. Sample applications that illustrate the possibilities on real examples can be very important tool in this task and can potentially inspire the future users.

The second possible, but probably never thoroughly implemented approach is based on the *user-centred design* paradigm (Hackos and Redish, 1998). It can be called a *top-down process*, as the starting point are complete research applications (or research tools) for the final users – H&SS scientists. Requirements for the applications should be discovered by applying methods of Context of Use Analysis. Next, research applications and the underlying network system of services and LT components should be designed and developed according to the requirements.

Despite the expected large number of LRTs that can be immediately re-used in the constructed infrastructure, this approach seem to be unrealistic. Research tools to be designed are innovative and are associated with the development of new research methods. Their discovery could be much easier through working prototypes and experiments for selected limited subdomains of H&SS. A long way from the project to the results and the perspective of costly long term investment could be unacceptable.

In comparison to the pure user-centred approach, a mixed option of a bi-directional process seems to be more practical. According to this approach, the existing LTRs, possibly many, are combined into a distributed network infrastructure, too. However, user-driven requirements are also taken into account. Designing the top level research applications for users is a starting point for many activities in the LTI development. The infrastructure construction process follows a metaphor of the Agile-like (Larman, 2004) light weight software designing method. Key users are identified and prototype research applications are created in co-operation with them and according to the requirements acquired from them. The application development stimulates the construction of technical fundaments, and inspires the identification of further user needs on the basis of analogy to the working prototypes.

4. Bi-directional approach of CLARIN-PL

In spite of significant improvement that had been made in the area of LT for Polish since 2005, quite many basic LTRs for Polish were still lacking at the start of CLARIN-PL (Jan. 2012). This situation resulted in deepening the technological barrier, as LRTS necessary for many applications simply did not exist. One of the most typical examples is the lack of a robust dependency parser for Polish – many application for English take the existence of such a parser for granted. Thus, the target CLARIN-PL structure is based on three parts:

1. CLARIN-PL Language Technology Centre[3] – the Polish node of the CLARIN distributed infrastructure,
2. a complete set of basic LRTs for Polish,
3. research applications for H&SS – first created for key users and selected H&SS sub-domains.

The LT centre is meant to provide fundamental CLARIN facilities (Roorda et al., 2009) like distributed authorisation and archiving system for LT supporting the CMDI meta-data format (Broeder et al., 2009) and persistent identifiers. A special focus is given to collecting LRTs for Polish and making them accessible via web services and linking them into processing chains. Moreover, the web services are accompanied by web-based applications with user interface in Polish[4]. A CLARIN centre with such functionality is classified as a CLARIN B-type centre (Roorda et al., 2009). As the number of resources (both text and speech) is limited among the CLARIN-PL partners, we plan to build interfaces linking the LT centre with the existing archives and repositories, e.g. digital libraries, and with other research infrastructures, e.g. DARIAH. However, the ongoing process of distributing the workload inside CLARIN ERIC causes that the Polish centre also plans

[2] www.clarin.eu

[3] http://www.clarin-pl.eu

[4] This requirement is very important, as many users from H&SS do not accept user interface in English.

to take responsibility for selected services that are fundamental for the whole infrastructure, i.e. elements of the responsibility of the CLARIN A-type centre.

A starting point for the identification of the missing LRTs for Polish was the comparison of the list of LRTs for Polish available on open licences with the BLARK[5] set of LRTs (Krauwer, 1998; Krauwer, 2003), as well as with the basic processing chains of Information Extraction. We envisaged the latter as the most likely scheme for applications. BLARK was selected as a *quasi* standard or standard *de facto* as the set of LRTs proposed in BLARK has already been a target point in development of LT for several languages. We assumed that implementation for Polish of all LRTs assumed in the BLARK set would increase interoperability between CLARIN-PL and the rest of CLARIN infrastructure, as the BLARK set is often used as a reference point. The bare existence of LRTs does not mean that they fit to CLARIN needs. Applications in H&SS impose high demands on their coverage and quality. Several of the existing LRTs must to be significantly expanded in CLARIN-PL in order to make them robust enough, see Sec. 5.

The multilinguality of CLARIN infrastructure makes the construction of bilingual resources crucial for interoperability. Balancing between the coverage and a range of resource types, we decided to concentrate mainly on bilingual Polish-English resources. An overview of LRTs planned to be developed in CLARIN-PL is presented in Sec. 5.

In a similar way to other CLARIN national consortia, general and flexible work-flows have to be constructed in order to facilitate the full use of different language tools.

Digital H&SS (also e-Humanities and e-Social Science) are developing very quickly, but they are still relatively new domains with not many fixed procedures. Research tasks in these domains are approached in a very dynamic way with a rich variety of specific solutions based on decisions dependent on research data and interim results. There are several methods for gathering informations about the Context of Use proposed in Human Computer Interaction (Hackos and Redish, 1998), but *direct observation* is considered to be the best one as it allows for observing users performing their tasks in their natural environment. In our case, the users are researchers and they perform scientific research. Users participating in the observation sessions should be representative. However, if the designer does not possess deep knowledge about the given domain, which still seems to be the case of Digital H&SS, the first group of the users, called *key users*, can be composed from those who are somehow characteristic to the domain. As key users, we selected scientists from H&SS who have already started using digital language-based methods in their research or are interested in applying LT in their research.

Direct observation is mainly based on collaboration with key users in their research projects. After collecting information about the Context of Use, possible LT-based techniques that can support the research are selected or even a new research process is defined. The method consists of several steps:

1. Establishing contacts with users
2. Identification of key users
3. Context of Use Analysis: users, their tasks and environments
4. Identification of the key applications corresponding to these users' tasks that can be supported by the available LT.

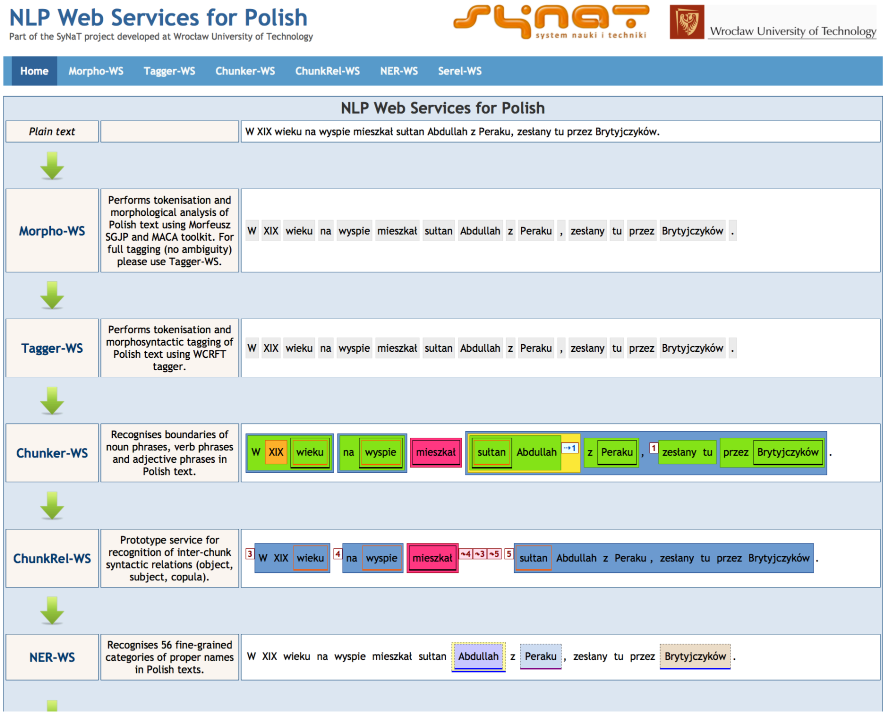
The first contacts with the prospective key users were established on the basis of the previous personal acquaintance with particular H&SS researchers. These direct links resulted in our participation in a couple of H&SS conferences and further contacts. Direct communication with possibly many conference participants appeared to be fruitful. Most key users are researchers who have already started using or are interested in using computer system in their research.

From the very beginning of the project we have been using our CLARIN-PL web page to inform potential users about the project. We published a list of of generally described potential CLARIN LTI applications with a special focus given to Polish LT. Our intention was to make H&SS researchers aware about existing possibilities and also to associate the CLARIN-PL web page with potential topics searched on the Web. We try to keep the list constantly growing, e.g. on the basis of the experiences collected during the application development. Moreover, on the CLARIN-PL web page portal we have also started collecting information about Polish conferences and projects from the domain of Digital H&SS and related domains – valuable information is the best advertisement on the Web.

The established contacts with the prospective users allowed us to select the first group of key applications discussed in Sec. 6. We tried to cover a maximal variety of research areas, but also to co-operate first with the most active users. During this first round, the number of applications is a less important factor, and we had aimed at only a few, e.g. due to the financial limitations. The available LT for Polish was also a limiting factor in this selection. We assumed that the first constructed applications would significantly broaden our understanding of the domain and help to identify further application domains or even generalise the constructed applications to general frameworks.

As a result, CLARIN-PL can be treated as an exemplar of a bi-directional approach combining together bottom-up and top-dow development. We are trying to harmonise these two approaches, i.e. to interactively shape the LRT development plan according to requirements collected from the work on key applications, e.g. CLARIN-PL tasks in the area of Information Extraction have been re-organised and re-ranked due to the collected experience. Summing up, the bi-directional approach is a fruitful scenario: key users,

[5] BLARK is the acronym for *The Basic Language Resource Kit* and it is "the minimal set of language resources that is necessary to do any precompetitive research and education at all" (Krauwer, 2003). A BLARK comprises different kind of resources and tools treated as a minimal required set for every language. This quasi-standard has been implemented for several languages and become a main reference point for evaluation of the state-of-the-art of the LT for a particular language.


NLP Web Services for Polish
Part of the SyNaT project developed at Wrocław University of Technology
system nauki i techniki
Wrocław University of Technology
Home
Morpho-WS
Tagger-WS
Chunker-WS
ChunkRel-WS
NER-WS
Serel-WS

NLP Web Services for Polish

Plain text		W XIX wieku na wyspie mieszkał sułtan Abdullah z Peraku, zesłany tu przez Brytyjczyków.
Morpho-WS	Performs tokenisation and morphological analysis of Polish text using Morfeusz SGJP and MACA toolkit. For full tagging (no ambiguity) please use Tagger-WS.	W XIX wieku na wyspie mieszkał sułtan Abdullah z Peraku , zesłany tu przez Brytyjczyków .
Tagger-WS	Performs tokenisation and morphosyntactic tagging of Polish text using WCRFT tagger.	W XIX wieku na wyspie mieszkał sułtan Abdullah z Peraku , zesłany tu przez Brytyjczyków .
Chunker-WS	Recognises boundaries of noun phrases, verb phrases and adjective phrases in Polish text.	W XIX wieku na wyspie mieszkał sułtan Abdullah →1 z Peraku , 1 zesłany tu przez Brytyjczyków .
ChunkRel-WS	Prototype service for recognition of inter-chunk syntactic relations (object, subject, copula).	3 W XIX wieku 4 na wyspie mieszkał ~4 ~3 ~5 5 sułtan Abdullah z Peraku , zesłany tu przez Brytyjczyków .
NER-WS	Recognises 56 fine-grained categories of proper names in Polish texts.	W XIX wieku na wyspie mieszkał sułtan Abdullah z Peraku , zesłany tu przez Brytyjczyków .

Figure 1: A prototype of the CLARIN-PL basic processing chain for Polish.

research tasks identified, key applications can provide generalisations resulting in adaptable research workflows.

5. Resources and Tools for Polish

The necessity of the significant improvement of basic LTRs for Polish was identified as a pre-requisite for lowering the technological barrier. On the basis of the analysis of the state-of-the-art of LT for Polish done at the start of CLARIN-PL, we planned several of tasks within this goal.

5.1. Resources and Supporting Technology

Concerning language resources, the starting point of CLARIN-PL was relatively good as several basic resources had been constructed and become matured, e.g. a huge National Corpus of Polish (Przepiórkowski et al., 2012), very large Polish wordnet – plWordNet (Maziarz et al., 2013) and an open KPWr corpus of Polish[6] with rich annotation (Broda et al., 2012). Thus, our main goals are completing the construction of selected resources and building bi-lingual resources and specialised corpora facilitating the envisaged needs of H&SS.

We plan to expand plWordNet to a comprehensive description of the Polish lexico-semantic system (with around 260 000 lexical units) and fully map it to Princeton WordNet 3.1 (Fellbaum, 1998)[7]. A large lexicon of the Multi-word Expressions described with the minimal constraints on their lexico-syntactic structures (Kurc et al., 2012) will be expanded up to the size of 60 000 Polish Multi-word Expressions manually described. All of them will be semantically described in plWordNet 3.0. The lexicon of semantically classified proper names (*NELexicon*[8]) will be expanded to 2.5 million distinct PNs. For both lexicons we will construct robust tools for their further automated expansion on the basis of corpora. This is meant to be an implementation of the idea of a dynamic lexicon, i.e. a combination of the core described manually and a large part extracted automatically from selected corpora on demand of the user. The automated tools will allow the CLARIN users to create their own domain extensions of both lexicons. The users will be also equipped in editors for the manual verification of the automatically extracted data. A large semantic valency lexicon for Polish predicative lexical units (verbs, nouns) will be also constructed (Hajnicz, 2014). Semantic restrictions on valency frame arguments will be described be described by means of the selected plWordNet synsets that are more general and define hypernymy subhierarchies used as represent ions of semantic domains.

Concerning corpora, CLARIN-PL is going to build: a transcribed training-testing Polish speech corpus, a corpus of Polish conversational texts transcribed from speech recordings and annotated parallel corpora mapping Polish text to several languages (Bulgarian, Russian and Lithuanian). Historical Polish corpus of text news from 1945-1954 that will be also developed is a resource directly fo-

[6] *Korpus Politechniki Wrocławskiej* (Wrocław Univeristy of Technology Corpus, `http://nlp.pwr.wroc.pl/kpwr`) in an open corpus of Polish which is balanced according to different genres and built from texts on Creative Commons. Currently, KPWr includes 449 000 tokens of text documents of 5 styles. KPWr has been annotated on several different levels of the linguistic structure, e.g. shallow syntactic structures (161 716 chunk annotations and), proper names, anaphora, semantic relations etc.

[7] As WordNet 3.1 appeared to be too small for providing mapping targets for all Polish senses we have initiated a significant expansion of WordNet 3.1 as a part of the CLARIN-PL plan.

[8] `http://nlp.pwr.wroc.pl/nelexicon`

cused on applications in H&SS.

In order to fully utilise the rich set of corpora, several systems for searching text and speech corpora will be expanded or built. A system for semantic indexing of large text corpora on the basis of publicly available encyclopaedias will be built.

5.2. Tools

The situation of the basic processing chain for Polish at the beginning of CLARIN-PL is presented below. Robust tools are presented in bold. Tools existing in prototypes with limited accuracy and coverage are written in normal font, and non-existing tools are shown in italic font.

1. **Segmentation into tokens and sentences.**
2. **Morphological analysis.**
3. Morphological guessing of unknown words (both without context and context sensitive).
4. **Morpho-syntactic tagging**.
5. Word Sense Disambiguation.
6. Chunker and shallow syntactic parser.
7. Named Entity Recognition and disambiguation.
8. Co-reference and anaphora resolution.
9. Temporal expression recognition.
10. Semantic relation recognition.
11. Event recognition.
12. *Shallow semantic parser.*
13. Deep syntactic parser *with disambiguated output*: dependency and *constituent*.
14. *Deep semantic parser.*

As most Polish language tools have been constructed with the focus on standard language and error-free text, an important element of the plan is the construction of a generic set of morpho-syntactic tools for Polish that can be adapted to a domain specified by the user.

We also plan to work on tools for the extraction of the semantic-pragmatic information from documents and collections of documents (e.g. keywords, semantic relations between text fragments and text summaries) and an open stylometric and textometric system.

All language tools presented above are used by CLARIN-PL or will be expanded or developed by CLARIN-PL. We plan to provide web services for all of them and also to include them into the processing chain. By now, we have implemented web services for: segmentation, morphological analysis, tagging, chunker and Named Entity Recognition[9]. Prototype web services for Word Sense Disambiguation and Semantic relation recognition are ready, but their accuracy is not yet satisfactory. There is also a web service providing access to plWordNet 2.2. Web services are accessible via both REST and SOAP and their programming interface is specified in WSDL language. We plan to describe them in CMDI meta-data format and integrate with the repository system of CLARIN-PL Language Technology Centre.

A prototype of the user interface for an implementation of the basic processing chain is presented in Fig. 1. The whole chain and its components are available as web services described with meta-data in CLARIN CMDI format. We plan to link them to WebLicht platform (Hinrichs et al., 2010) and also to build our own platform for defining processing chains focused on Polish users.

6. Research Applications

Only programs or systems constructed in response to real CLARIN users' (i.e. H&SS researchers) requests can be treated as CLARIN applications. Interactive systems that are not used do not exist.

A search system for the corpus of conversational data *Spokes*[10] was constructed in the close co-operation with linguists inside the CLARIN-PL consortium. So, it is not a genuine application, but it provides rich facilities for not only searching the corpus, but also for statistical analysis of the retrieved data. Corpus search tools are basic application that mostly provide only searching through language data, but anyway they are crucial applications. However, the issues of rich annotation, big data volumes and statistical analysis of the query results, the construction of the corpus search tools is much more challenging.

Requests from users sometimes reveal gaps in the available technology that were not expected before the project start. Several tools for web-based corpus building appeared to be too sensitive to text encoding errors found in the web (e.g. a different code page declared in meta-data than really used). As a result a system for collecting Polish text corpora from the Web had to be constructed. The system is combined with morphological analysis in order to detect texts including larger number of errors (or non-words). The system was also requested to provide support for semi-automated extraction from blogs only those elements that fit to the pre-defined user requirements.

There were several textometric and stylometric tools available, but none of them was well suited for rich inflection of Polish, e.g. the available tools did not provide support for lemmatisation and tagging of Polish. We plan to build a system for Polish enabling the use of features defined on any level of the linguistic structure: from the level of word forms up to the level of the semantic-pragmatic structures. The system will combine several existing components: language tools for pre-processing, *Fextor* (Broda et al., 2013) – a system for defining features in a flexible way, *Stylo*[11] – a stylometric package for English, *SuperMatrix* (Broda and Piasecki, 2013) – a system for building and processing very large co-incidence matrices with linking to clustering and Machine Learning packages.

[9] The services are available at `www.clarin-pl.eu/en/services/`

[10] `http://clarin.pelcra.pl/Spokes/`

[11] `http://crantastic.org/packages/stylo/versions/34587`

Stylometric techniques appear to be applicable in many tasks of H&SS that are based on the comparison of texts, e.g. in sociology (features that are characteristic for different subgroups), political studies (similarity and differences between political parties), literary studies (analysis of blogs as creative work types), etc. The extended system will allow for the analysis of the semantic associations of words on the basis of Distributional Semantics, semantic relation extraction, collocations, semantic comparison of texts and texts collections, etc. Thus, it starts to be similar to a system for the semantic text classification discussed in the next subsection.

6.1. Semantic Text Classification for Sociology

One of our first research application is an exemplar of the scheme which can be generalised to many projects in H&SS. One of the goals of the research project realised in *Collegium Civitas* (a non-state university) in Warsaw was to check the content of web pages of the Polish institutions (public and private) related to culture, in its broadest sense. Around 3200 institution were pre-selected and almost 200 000 documents were acquired from their web sites. The content of the web pages was divided into paragraphs of different sizes (around 1 200 000). The goal was to classify the paragraphs into 20 semantic classes defined by the sociologists. The classes describe different aspects of the use of the web page as a communication medium and they were organised into three groups: competences, functions of the culture, thematic areas plus 6 individual classes (e.g. auto-presentation or local function).

The initial vision was a simple system for supervised classification of text documents. After the Context of Use Analysis, the plan was expanded to a complex system encompassing user-controlled corpus building, text preprocessing (text segmentation and morpho-syntactic tagging and parsing), automated sample selection, manual annotation, training classifiers and automated annotation and result analysis. Moreover, we discovered that there is no open corpus annotation editor focused on applications in Social Sciences. The constructed prototype system can be also adopted to many similar tasks in Digital H&SS.

6.2. Literary Map

Literary Map is a CLARIN-PL application that has originated from a concrete user request formulated during an open part of a CLARIN-PL working meeting. The first prototype is presented in Fig. 2. The user is Digital Humanities Centre of The Institute of Literary Research of PAS. The main idea is to identify all geographical names in the literary text (or a corpus) and map them onto the geographical map. The task goes beyond Named Entity Recognition (NER), as NER must be combined with geo-location. We use geo-location service provided by Google, but still location PNs recognised in text must be grouped into expression recognised by Google in a way enabling good accuracy of locating them. We proposed to expand the initial idea with recognition of semantic relations linking non-spational PNs in the text with the location PNs and visualising those links on the map, too. Recognition of the temporal expression could further enrich the application.

Two scenarios of use are considered: fully automated and bootstrapping. According to the first, users process whole corpora of literary texts and next can analyse collected statistical data or browse mapping of the individual texts. However, due to the limited accuracy of the whole system, the second scenarios in which the system is used a supporting tool during corpus annotation with mappings is more likely in research – annotations proposed by the system are next corrected by the researchers.

7. Conclusions

The Quality of Service notion is very rarely used in relation to LRTs and LTI. However, this is the crucial question: for what research tasks and what scenarios are our LRTs good enough? If we aim at fully automated procedures, the expected quality is very high, e.g. 5% can bias a lot statistical analysis of data extracted from a corpus. Application of LT to the research in H&SS seem to be much more challenging than in commercial systems! We need to develop a model of LT-based applications in which we can describe and manage errors introduced by different LRTs and their accumulated influence on the final result of the whole application.

Semi-automated model in which LT-based applications are used for preparing initial text annotation, next corrected by researchers, or supporting researchers in browsing corpora and finding examples is the most likely way. Here, Visualisation of the results on different stages comes into play as a very important element of LTI.

Any model of LTI we aim for users should be the starting point for LTI development and also the goal for this work.

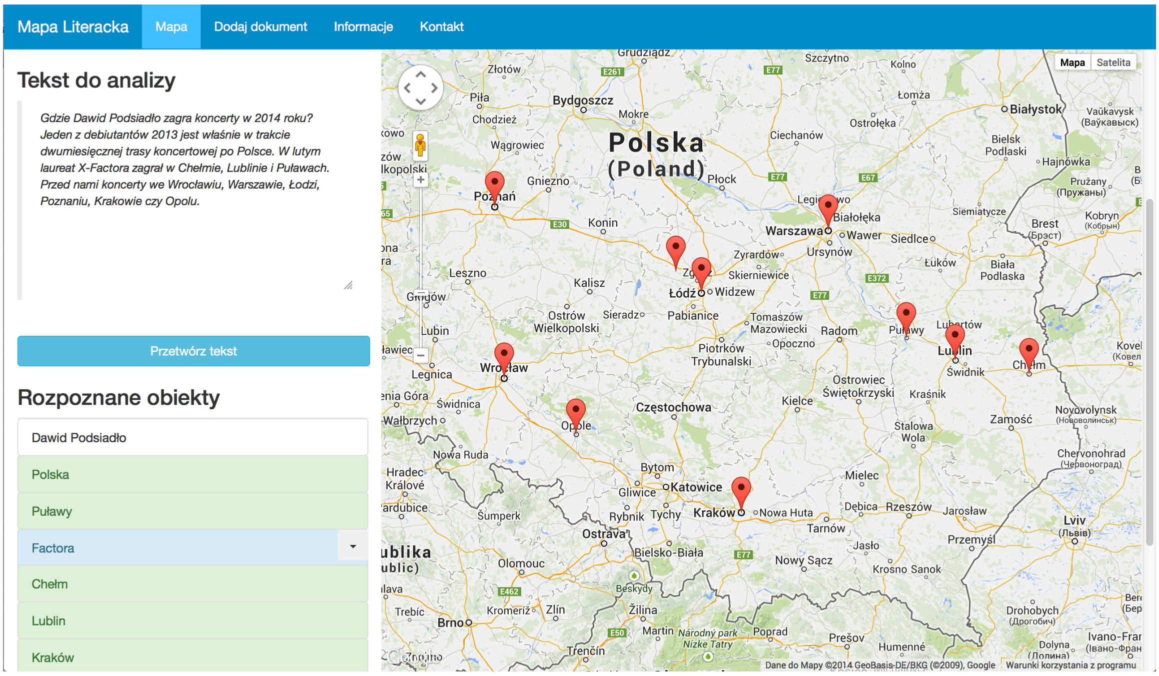
Acknowledgement: The work was financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and was supported by the EU's 7FP under grant agreement no 316097 [ENGINE].

8. References

Bartosz Broda and Maciej Piasecki. 2013. Parallel, massive processing in SuperMatrix – a general tool for distributional semantic analysis of corpora. *International Journal of Data Mining, Modelling and Management*, 5(1):1–19.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In N. Calzolari et al., editor, *Proc. of the Inter. Conference on Language Resources and Evaluation, LREC'12*, Turkey. ELRA.

Bartosz Broda, Paweł Kędzia, Michał Marcińczuk, Adam Radziszewski, Radosław Ramocki, and Adam Wardyński. 2013. Fextor: A Feature Extraction Framework for Natural Language Processing: A Case Study in Word Sense Disambiguation, Relation Recognition and Anaphora Resolution. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem, and Piotr Fuglewicz, editors, *Computational Linguistics*, volume 458 of *Studies in Computational Intelligence*, pages 41–62. Springer Berlin Heidelberg.



Tekst do analizy

Gdzie Dawid Podsiadło zagra koncerty w 2014 roku? Jeden z debiutantów 2013 jest właśnie w trakcie dwumiesięcznej trasy koncertowej po Polsce. W lutym laureat X-Factora zagrał w Chełmie, Lublinie i Puławach. Przed nami koncerty we Wrocławiu, Warszawie, Łodzi, Poznaniu, Krakowie czy Opolu.

Przetwórz tekst

Rozpoznane obiekty

- Dawid Podsiadło
- Polska
- Puławy
- Factora
- Chełm
- Lublin
- Kraków

Figure 2: A prototype of the Literary Map application.

Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Erhard Hinrichs, Lothar Lemnitzer, Dieter van Uytvanck, Andreas Witt, and Peter Wittenburg. 2009. Registry requirements metadata infrastructure for language resources and technology. Technical Report CLARIN-2008-5, Consortium CLARIN. PID: `http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-33`.

Christiane Fellbaum, editor. 1998. *WordNet — An Electronic Lexical Database*. The MIT Press.

J. Hackos and J. Redish. 1998. *User and Task Analysis for Interface Design*. Wiley Comp. Pub.

Elżbieta Hajnicz. 2014. Lexico-semantic annotation of *składnica* treebank by means of PLWN lexical units. In Orav et al. (Orav et al., 2014), pages 23–31.

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based lrt services for german. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 25–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

ISO. 1997–1999. *ISO 9241 — Ergonomic Requirements for Office Work with Visual Display Terminals*. ISO.

Steven Krauwer. 1998. BLARK: The Basic Language Resource Kit. ELSNET and ELRA: Common past, common future. Web page. Accessed 16th Sep. 2014.

Steven Krauwer. 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM 2003*.

Roman Kurc, Maciej Piasecki, and Bartosz Broda. 2012. Constraint based description of polish multiword expressions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2408–2413, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Craig Larman. 2004. *Agile and Iterative Development: A Manager's Guide*. Addison-Wesley.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In *Proc. of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452, Hissar, Bulgaria. INCOMA Ltd.

Heili Orav, Christiane Fellbaum, and Piek Vossen, editors. 2014. *Proceedings of the 7th International WordNet Conference (GWC 2014)*, Tartu, Estonia. University of Tartu.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Dirk Roorda, Dieter van Uytvanck, Peter Wittenburg, and Martin Wynne. 2009. Centres network formation. Technical report CLARIN-2008-3, Consortium CLARIN. PID: `http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-27`.

Peter Wittenburg, Núria Bel, Lars Borin, Gerhard Budin, Nicoletta Calzolari, Eva Hajicová, Kimmo Koskenniemi, Lothar Lemnitzer, Bente Maegaard, Maciej Piasecki, Jean-Marie Pierrel, Stelios Piperidis, Inguna Skadina, Dan Tufis, Remco van Veenendaal, Tamás Váradi, and Martin Wynne. 2010. Resource and service centres as the backbone for a sustainable service infrastructure. In N. Calzolari et al., editor, *Proc. of the International Conference on Language Resources and Evaluation, LREC 2010, Malta*, pages 60–63. ELRA.