Training Workshops in the Bi-directional Model of the Language Technology Infrastructure Development

Maciej Piasecki
maciej.piasecki@pwr.edu.pl

Jan Wieczorek
jan.wieczorek@pwr.edu.pl

Wroclaw University of Science and Technology
Department of Computational Intelligence
G4.19 Research Group

Wrocław University
of Science and Technology

Department of
Computational
Intelligence

Users as a basis for Language Technology Infrastructure

Interrelations between users and LTI:

- User needs as a reason for LTI
- Addressing user needs as a fundamental challenge
- LTI must have users to gain funding

Users as a basis for Language Technology Infrastructure

CLARIN-PL LTI users - description:

- researchers and students...
- ...from various Social Sciences & Humanities disciplines...
- ...in need of some data processing chains useful for analysing language data...
- ...that are customized to the nature of their research needs

- LT users are significantly different from LT developers
- LT users are a reason for a LTI to exist

Schemes of LT development

- Bi-directional development of LTI (Piasecki, 2014)
- Bottom-up - development of the necessary basic elements of LTI
  - a distributed network infrastructure
  - basic LT processing chain
- Top-down
  - user's needs web-based research applications
  - close co-operation with key users from the H&SS domain
  - amendments to the shape of the technical basis: LRTs, standards,
  - inspirations, identification of the further user needs, next iterations …

Schemes of LT development

Top-down part: process

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 Language Technology

Schemes of LT development

CLARIN-PL bi-directional development of LTI – examples of applications

- LEM: http://ws.clarin-pl.eu/lem.shtml
  - Literary Eploration Machine – statistical analysis of texts
- WebSty: http://ws.clarin-pl.eu/websty.shtml
  - A open textometric and stylometric system focused on Polish, and selected other languages (WebStyML)
- Inforex – web-based editing and annotation of corpora
- WordnetLoom – visual browsing and editing of semantic lexicons
- Spokes – a search engine for speech corpora
- Paralella – a search engine for parallel corpora
- ChronoPress – browsing and statistical analysis of historical corpora

Ways to reach users

Popular methods of dissemination:

- advertisements,
- on-line and printed materials,
- publications,
- lectures

Main disadvantages of these methods:

- focusing on promotion of finished products
- communication works mainly in one direction - from developers to potential users

Ways to reach users

CLARIN-PL prefers workshops as a method of dissemination because of:

- live contact
- users' feedback

These are the key aspects of bi-directional LTI's development model.

Ways to reach users

Types of workshops used by CLARIN-PL:

General Workshops – overview presentation of the current state of CLARIN-PL infrastructure, aimed at broad scientific audience (e.g. subsequent editions of training workshops “CLARIN-PL in Research Practice”)

Targeted Workshops – focused on selected subdomains, with specified types of research tasks, sometimes aimed at users’ subgroups or subcommunities, organised for smaller number of participants

User-defined Workshops – mostly initiated by users, sometimes originating from their very specific needs (characteristics of the field of study), sometimes from curiosity

General Workshop

In 2015 CLARIN-PL started to organise General Workshops

During the first three editions we have presented:

- prototypes of tools and applications designed for SS&H researchers
- prototypes or projects of language resources

One could say, that it was (at the time) perhaps to early – in terms of tools and resources development – to organise such an event.
But our objective was clear.

General Workshop

What did we try to achieve?

- to increase the awareness of possible LT applications
- to have a broad look for potential active users (researchers) of CLARIN-PL LTI
- to recognise potential users' needs (which was very hard to achieve due to their lack of awareness)

Feedback obtained during first two editions of General Workshops was crucial in the evolution of CLARIN-PL Infrastructure:

- we have acquired people testing the efficiency of our prototypes
- we were able to accurately identify the real needs of researchers
- in the result tools/resources can now optimally meet the demand
- we were able to plan the next steps in the development of LTI

General Workshop

Indirect effects of organising General Workshops:

- stabilisation of CLARIN-PL funding – every year The Ministry of Science and Higher Education receives a report on new research carried out using CLARIN-PL LTI and it ensures the ministry to continue financing the project. Large number of new users recruits from former workshop participants or are the researchers from teams, where workshop participants work.
- extension of the workshop offer by two other types. Most of the initiators/participants of Targeted Workshops and User-defined Workshops are former General Workshops participants or their collaborators.

General Workshop - example

10th edition of our General Workshop: CLARIN-PL in Research Practice
Date: 25 - 26 IX 2019
Place: Maria Skłodowska Curie University in Lublin, Faculty of Political Science
Number of participants: 60

General Workshop - example

Programme (a good example of a typical GW programme):

	Room no. 1	Room no. 2
	Wednesday (25 IX)	
09:15	CLARIN-PL: Introduction (M. Piasecki)	
10:15	The Polish Valence Dictionary Walenty (E. Hajnicz)	Polish Parliamentary Corpus (M. Ogrodniczuk)
11:30	Coffee break	
12:00	CLARIN-PL lexical-semantic resources: plWordNet with its extensions and Lexical Platform (A. Dziob)	Tools for creating corpora: DSpace, Clarin Cloud, Inforex (M. Marcińczuk, J. Wieczorek)
14:00	lunch	
15:00	Chronopress: a corpus of selected Polish press texts from 1945-1963 (A. Pawłowski)	Morphological Analyzer Morfeusz (K. Krasnowska-Kieraś)
16:15	Speech Processing Tools (M. Kleć)	Dependency parsing (A. Wróblewska)
17:45	Individual consultations	
20:00	Workshop dinner	

	Room no. 1	Room no. 2
	Thursday (26 IX)	
09:00		Literary Exploration Machine (LEM) & Simple Text Statistics (T. Walkowiak, M. Piasecki)
10:30	A tool for creating corpora: Korpusomat (Ł. Kobyliński)	A tool for extracting terminology from texts - TermoPL (P. Rychlik)
11:45	Coffee break	
12:15	Tools for creating corpora: DSpace, Clarin Cloud, Inforex (M. Oleksy, J. Wieczorek)	A tool for texts similarity analysis - WebSty & information extraction from text (T. Walkowiak, M. Piasecki)
14:15	lunch	
15:15	Multilingual Corpora (mainly Slavic and Baltic) (R. Roszko)	
16:15	Corpora browsers: Kontext + Korpusomat (J. Wieczorek, M. Oleksy)	Individual consultations
17:45	Summary and closure (M. Piasecki)	
18:00	The end	

General Workshop - example

Why Lublin?

- discussions with participants previous editions of GW
- it was repeatedly pointed out that we should conduct more promotional activities in Social Sciences.
- The interest in this community in the years 2015-2018 was very low. However, with time it started to grow.
- Therefore, we decided to find a university partner among Social Sciences faculties in the part of Poland, where we had not yet organized workshops (south-east).

Besides...
Lublin is a beautiful city!

General Workshops - summary

Facts:

- ten editions between 2015 and 2019
- over 750 participants
- reduced the frequency of GW down to 1 per year (in 2018 and 2019)

Positive effects:

- awareness about CLARIN-PL and Polish LI in SS&H research increased significantly
- established cooperation with several **key users** (who are also very supportive in the development of tools) and many contacts across different SS&H areas
- as a snowball effect, new invitations to organise workshops are coming

General Workshop - summary

Downsides:

- only a small number of users stays active after a workshop
- time spent at workshops is not sufficient to discuss different issues (or maybe: topics are too diversified)
- the audience seems to be too diversified according to their background, skills and research tasks
- large total number of participants and limited time result in groups of 20-30 participants during practical classes; groups are to big to be manageable with respect to individual help
- the impact on the uptake of CLARIN-PL in research projects was also smaller than expected

Targeted Workshop

Goal: to improve the participants' competence and to train them up to the level of working alone with CLARIN-PL resources and services.

Users enrolled for the workshop should share similar disciplines, research objectives or should share a need of the same language data.
Participants shall be experienced in their fields, but are not expected to have skills in LT.

Workshop groups shall not exceed 24 persons.

An initiative to organise a workshop is based on cooperation between users and CLARIN-PL. It may originate from prior support for the users' research tasks or contacts established during other workshops.

Classes conducted during the targeted workshops are strictly focused thematically. They usually concern a coherent set of issues.

Targeted Workshop - examples

February 2017: A discourse researcher's team (from University of Silesia) participated in a general workshop in Łódź.

April 2017: They invited CLARIN-PL to provide a targeted workshop on the use of language technology in diachronic linguistics during a conference on Polish diachronic corpora (Katowice).

June 2019: CLARIN-PL provided a series of consultations and training organised by silesian team.

Autumn 2019: Now this research team is going to submit three new applications for research grants:

1. study of the discourse of memories in the archives of oral history;
2. development a historical corpus of Polish theatrical dramas;
3. establishing a platform for integration of dispersed historical corpora of the Polish language.

Targeted Workshop - summary

Facts:

- six editions of targeted workshops between 2017 and 2019
- over 100 participants
- gradually increasing frequency up to 2-3 per year (in 2018 and 2019)

Positive effects:

- testing new functions of the tools and collecting new requirements
- significant increase in the participants' competences and users’ activity in the CLARIN-PL Centre

Downside:

- this type of workshop involves the CLARIN-PL Centre's employees very much; it requires close cooperation with the initiators and substantive preparation (not only in the field of NLP)

User-defined Workshop

Main goal and characteristic of users are similar to those of Targeted Workshops.

Workshop groups shouldn't exceed 20 people. Training groups are usually much smaller.

The initiative to organise a workshop comes from external partners, who also have a significant influence on the programme. (the main difference between TW and UdW)

Usually the organizers of this type of events are scientific institutions, to which users belong. In some cases, researchers initiating workshops contact CLARIN-PL on the basis of a recommendation from another CLARIN ERIC partner or other Polish researcher, already experienced with LT in SS&H.

Practical classes revolve around a narrow topic. The training itself is usually very proactive: research tasks are solved on the basis of material previously submitted by the organizers, and participants are also already familiar with it.

User-defined Workshop - examples

- EngHum workshops in Warsaw - November 2018, LTI offer for researchers of endangered languages

- Workshops conducted for the needs of SADiLAR on building the African Wordnet (Pretoria, February 2019)

- Training on the use of the CLARIN-PL repository (DSpace) and the Kontext corpus search engine for employees of the Institute of World Art Studies (Warsaw, January 2019)

- training on the use of the WebSty and Lem applications for the discourse - researchers from Adam Mickiewicz University (Poznań, March 2019)

Workshop Organisation Procedure

A) Initiative and contact:

General Workshops:

- the initiative is always taken by CLARIN-PL, which defines the thematic scope of the programme and selects a convenient place.
- Next, we are looking for a partner who has appropriate infrastructure (rooms, Internet access, etc.) and is potentially interested in the subject of the workshops.

Targeted workshops and user-defined workshops:

- organisation of targeted and user-defined workshops usually results from the previous contacts, e.g. scientific events with presentation of research done with CLARIN-PL tools or events during which CLARIN-PL tools were presented or recommended, or authors mentioning CLARIN, or a quotation in an article or book

Workshop Organisation Procedure

B) Identifying the needs of participants:

- the programme of the workshop, the type of activities, the way they are conducted is strictly determined by the needs and expectations of future participants.
- General Workoshops: the organisers assume low LT competence of the participants and also lack of their specific expectations
- TW & UdW: The organisation of other types of workshops requires an interview with experts (partners) who know the needs and competences of a specific (potential) participants. This interview takes the form of a dialogue in which partners from SS&H describe their needs and the CLARIN-PL staff selects the appropriate tools/resources that may suit these criteria.

Workshop Organisation Procedure

C) Setting up the programme:

- the programme of a targeted and user-defined workshop is always based on the identified and defined needs of the participants
- the programme is focused on one or two thematic threads (e.g., information extraction from texts and tools supporting lexicographic work or tools supporting work of translators and creation/annotation of language corpora)
- during the general workshop all activities are focused on demonstrating the tools' functionality, their possible applications and use examples

Thank you!