

REPORT

ENeL TRAINING SCHOOL 2016

Tools and methods for creating innovative e-dictionaries

Central Hotel, Ljubljana, Slovenia

17 – 20 May 2016, Ljubljana, Slovenia

Rute Costa
Training Manager



1. Introduction	3
1.2. Objectives of WG3: INNOVATIVE E-DICTIONARIES	3
2. Objectives and Program of the Training School	4
2.1. Objectives of the training school	4
2.2. Schedule	5
2.3. Methods, Class Materials and Software	6
3. Training Team	6
4. Trainees	7
4.1. Selected Candidates	7
5. Facilities	11
5.1. HOST	11
5.2. Accommodation	11
6. Budget	12
Annex 1: Certificate of participation	13
Annex 2: Attendance List	14

1. Introduction

ENeL, the European Network of e-Lexicography has been established to increase, coordinate and harmonise European research in the field of e-lexicography and to make authoritative information on the languages of Europe easily accessible.

ENeL is subdivided in 4 Working Groups:

WG1 Integrated interface to European dictionary content;

WG2 Retro-digitized dictionaries

WG3 Innovative e-dictionaries

WG4 Lexicography and lexicology from a pan-European perspective.

One of the main aims of ENeL is to organise a series of three Training Schools designed to underpin the activities of WG 2, WG3 and WG4, whereas WG1 will get most of its scientific input from these three WGs.

The first ENeL Training School was held in Lisbon, from the 6th to the 10th of July, hosted by the Faculdade de Ciências Sociais e Humanas of the Universidade Nova de Lisboa to reinforce the activities of the WG2.

The second ENeL Training School was held in Ljubljana, from the 17th to the 20th of May, hosted by the University of Ljubljana (Centre for Language Resources and Technologies) and Trojina (Centre for Applied Linguistics) to reinforce the activities of the WG3.

1.2. Objectives of WG3: INNOVATIVE E-DICTIONARIES

The third working group will focus on the development of digitally born dictionaries, focusing on the latest developments in e-lexicography and the interface between lexicography and computational linguistics. Work will be carried out on:

- the description of the workflow for corpus-based lexicography
- providing an overview of existing software needed to set up this workflow
- dictionary writing systems

- the analysis of the possible impact of automatic acquisition of lexical data
- the analysis of the interface between dictionary and computational lexica (cf. wordnets) and syntactically and semantically annotated corpora (cf. FrameNet, SemCor, Senseval)
- the investigation of the possible use of dictionary content for computational linguistic applications
- organising a training school on innovative approaches in e-lexicography in 2016.

2. Objectives and Program of the Training School

2.1. Objectives and topics of the training school

In the context of ENeL Working Group 3 innovative dictionaries are defined as digitally-born. This means that they no longer resemble traditional paper dictionaries but try to fully exploit the new possibilities of the digital medium using methods from computational linguistics. In this Training School participants will learn about tools and methods that are used to create innovative e-dictionaries. Participants will learn how to collect and analyse (corpus) data for a chosen language, automatically extract data using one of the available extraction procedures or applications, and how to edit and publish extracted data in one of the available (online) dictionary writing systems and/or publishing platforms. The programme includes both theoretical courses and hands-on sessions.

Topics:

- Introduction (presentation of different successful innovative e-dictionaries projects)
- Corpus building (Corpus Query Systems, Web as Corpus)
- Data modelling, automatic extraction procedures
- Online publication and web design (Dictionary Writing Systems, online publication)

2.2. Schedule

	Tuesday, 17th	Wednesday, 18th	Thursday, 19th	Friday, 20th
	CORPUS BUILDING DAY	DATA MODELING / CQS DAY	EXTRACTION DAY	DWS DAY
9.00-9.30	Introduction (S. Krek and presenters)	Corpus building: results (E. Stemle)	Automatic extraction of data: introduction (S. Krek, C. Tiberius)	Presentation of DWS: overview (M. Měchura)
9.30- 10.00	Presentation of participants	Corpus building: annotation (E. Stemle, M. Jakubíček)	Automatic extraction of data: tools (S. Krek, C. Tiberius, M. Jakubíček)	Presentation of DWS (Lexonomy) (M. Měchura)
10.00- 10.30		Corpus building: Sketch Engine (M. Jakubíček)		
10.30- 11.00	Coffee break	Coffee break	Coffee break	Coffee break
11.00- 11.30	Presentation of participants	Data modelling: introduction (C. Tiberius)	Presentation: extraction (API) (M. Jakubíček, I. Kosem)	Practical session: setting up (M. Měchura)
11.30- 12.00	Presentation of projects: ANW (C. Tiberius)	Data modeling: Lexonomy (M. Měchura)		Practical session: importing data (M. Měchura)
12.00- 12.30	Presentation of projects: SLD (S. Krek)	Practical session: data modelling (C. Tiberius, M. Měchura)		Practical session: editing (M. Měchura)
12.30- 14.00	Lunch	Lunch	Lunch	Lunch
14.00- 14.30	Corpus building: introduction (E. Stemle)	Corpus Query Systems: Sketch Engine (C. Tiberius, M. Jakubíček)	Practical session: extraction (API) (M. Jakubíček, I. Kosem, S. Krek)	Presentation of DWS: publishing (M. Měchura)
14.30- 15.30	Corpus building (E. Stemle)	Corpus Query Systems: CQL (grammar) (C. Tiberius, M. Jakubíček) Practical session: grammar writing (C. Tiberius, M. Jakubíček)		Practical session: publishing (M. Měchura)
15.30- 16.00	Coffee break	Coffee break	Coffee break	Coffee break
16.00- 17.00	Practical corpus building session (E. Stemle)	Corpus Query Systems: GDEX (I. Kosem)	Practical session: extraction (advanced) (M. Jakubíček, I. Kosem, S. Krek)	Presentation of participants' results



2.3. Methods, class materials, software's

After the presentation of the trainees, the training school began with the presentation of the following projects:

- i. ANW (an online scholarly dictionary of contemporary standard Dutch)
- ii. SLD (Slovenian Lexical Database)
- iii. ELEXIS - European Lexicographic Infrastructure (Call: Integrating Activities for Starting Communities) – Project INFRAIA-02-2017

The presentation was followed by activities concerning Corpus building. For this purposes trainees were introduced to [www.sketchengine.co.uk/#blue] Sketch Engine and BootCat [<http://bootcat.sslmit.unibo.it/>]. The rest of the activities were dedicated to practical sessions.

After the results had been discussed in larger groups, the training followed with the demonstration of Lexonomy [www.lexonomy.eu], a Dictionary Writing System (DWS) tool. Then students could exercise and discuss the solutions undertaken and the results they achieved during the practical session. An introduction was made to Corpus Query Systems (CQS):

- i. CQL – Corpus Query Language
- ii. GDEX – Automatically finding good dictionary examples in a corpus

The following step was related with the automatic extraction of data. The trainees continued to work with Sketch engine. They learned to use API (application programming interface for lexical computing) extraction and edit Sketch grammar to extract advanced data with grammatical relations.

Finally the last day, trainees were invited to set-up an e-dictionary by importing data in XML format, from Sketch engine into Lexonomy. Each trainee had the opportunity to edit and publish their work. By the end of the day, every trainee get some time to make a short oral public presentation of the work and get feed-back from all the members of the group.

The whole information had been made available for all participants in Dropbox.

3. Training Team (alphabetical order)

The training team were constituted be the following trainers:

Erggon Stemle, EURAC, Italy

Miloš Jakubiček, LCL, Czech Republic

Simon Krek, University of Ljubljana, Slovenia

Iztok Kosem, Centre for Applied Linguistics, Trojina Institute, Slovenia

Carole Tiberius, Institute for Dutch Lexicology, the Netherlands

Michal Boleslav Měchura, Dublin City University and Foras na Gaeilge, Ireland

4. Trainees

On the ENeL website¹ the information were given on “who can apply” and “how to apply”. In February, over Skype, the team had a meeting to analyse the candidates applications (brief CV and motivation letter) and came up with the following results:

Number of candidate applications: 30

Selected trainees: 30

Note: 8 local trainees get no funding

4.1. Selected Candidates

List of the selected candidates:

Nr	Trainee	Institution	Country	Gender
1	Said Abafar	Universum ICT	NL	M
2	Kamil Barbierik	Czech Technical University	CZ	M
3	Kristina Bobekova	Slovak Academy of Sciences in Bratislava	SK	F
4	Milena Hebal-Jezierska	University of Warsaw	PL	F
5	Alenka Jelovsek	Institute of Slovenian Language	SI	
6	Margarida Ramos	Universidade Nova de Lisboa	PT	F
7	Kristina Koppel	Institute of the Estonian Language	EE	F
8	Martina Kramaric	Institute of Croatian Language and Linguistics	HR	F
9	David Lindemann	University of the Basque Country	ES	M
10	Marcas Mac Coinnigh	Queen's University Belfast	UK	M
11	Mija Michelizza	Fran Ramovš Institute of the Slovenian Language, Research Centre of the Slovenian Academy of Sciences and Arts	SI	F
12	Seanan O Coistin	University of Dublin	IE	M
13	Ildiko Pilan	University of Gothenburg	SE	F
14	Behrang Qasemizadeh	Heinrich-Heine-Universität Düsseldorf	DE	M

¹ <http://www.elexicography.eu/events/training-schools/ljubljana-2016/>

15	Ursula Schulze	Trier Center for Digital Humanities Trier University	DE	F
16	Cormac Breathnach	An Coiste Téarmaíochta	IE	M
17	Renate Bronikowska	Institute of Polish Language, Polish Academy of Sciences	PL	F
18	Katarina Chlpikova	Slovak Academy of Sciences in Bratislava	SK	F
19	Charles Dillon	Royal Irish Academy	UK	M
20	Helena Dobrovoljc	She didn't attend	SI	F
21	Kaja Dragoljevic	University of Ljubljana (student)	SI	F
22	Theodorus Franssen	University of Dublin	IE	M
23	Anna Havinga	University of Bristol	UK	F
24	Fahrettin Sirin	Freelance translator (PhD candidate: Bielefeld University)	DE	M
25	Marie Steffens	Université de Liège	BE	F
26	Tomislav Stojanov	Institute of Croatian Language and Linguistics	HR	M
27	Janos Jezovnik	Fran Ramovš Institute of the Slovenian Language, Research Centre of the Slovenian Academy of Sciences and Arts	SI	M
28	Urska Vranjek Oslak	Fran Ramovš Institute of the Slovenian Language, Research Centre of the Slovenian Academy of Sciences and Arts	SI	F
29	Tanara Zingano Kuhn	Universidade de Lisboa	PT	F
30	Maja Loncar	Institute for Croatian Language and Linguistics	HR	F

In conclusion: 29 trainees attended the Ljubljana Summer School:

20 Female / 9 Male

10 different countries: Portugal (1), Slovakia (2), Poland (6), Croatia (4), Spain (2), Greece (3), Serbia (3), Germany (5), Romania (1) and Ireland (1)

25 different institutions, as follows:



Nr	Institution	Country
1	Institute of Croatian Language and Linguistics Zagreb	Croatia
2	Christian-Albrechts-University Kiel	Germany
3	General and Comparative Linguistics University of Regensburg	Germany
4	Trier Center for Digital Humanities Trier University	Germany
5	Research Center for Modern Greek Dialects-ILNE Academy of Athens	Greece
6	Research Center of Scientific Terms and Neologisms, Academy of Athens	Greece
7	University of Amsterdam	Ireland
8	Adam Mickiewicz University Poznań	Poland
9	Department of Formal Linguistics University of Warsaw	Poland
10	Department of Medieval Latin, Institute of Polish Language Polish Academy of Science	Poland
11	Institute of Applied Polish Studies University of Warsaw	Poland
12	Institute of English Studies University of Warsaw	Poland
13	Institute of Southern and Western Slavonic Studies University of Warsaw	Poland
14	Faculty of Social Sciences and Humanities,	Portugal

	Universidade Nova de Lisbon	
15	Institute of Romanian Philology Romanian Academy, Bucharest	Romania
16	Institute for Serbian Language SASA Belgrade	Serbia
17	Ľ. Štúr Institute of Linguistics Bratislava	Slovakia
18	School of Philology at the Universidad Complutense of Madrid (UCM)	Spain
19	UPV-EHU University of the Basque Country	Spain
20	Bielefeld University	Germany
21	Serbian Language Institute SASA Belgrade	Serbia
22	Institute of Croatian Language and Linguistics	Croatia
23	Serbian Language Institute of SASA Belgrade	Serbia
24	Institute for Language and Speech Processing Athens	Greece
25	Christian-Albrechts-University University Kiel	Germany

5. Facilities

5.1. Host

Funded by COST Action IS 1305 the University of Ljubljana hosted the second 'European Network of e-Lexicography' Training School in Ljubljana.

The whole Training School took place in a room at the Central Hotel, in Ljubljana. The trainees used their own software; the requested software where accessible for free during the whole week

5.2. Accommodation

There are many good accommodation options in the centre of Ljubljana, we list a selection of them below, most of them within a walking distance of the workshop venue. For the best deals, please check websites such as Trivago, Booking.com or Hotels.com; also, always check the homepage of the hotel for special deals.

Hostels

- Hostel Celica: <http://www.hostelcelica.com/>
- Bed & Breakfast Atticus: <http://www.atticus.si/>
- Hostel DIC: <http://hosteldic.com/SL/>
- H20 Hostel: <http://www.h20hostel.com/>
- HOSTEL 24: <http://www.hostel24.si/>

Basic Hotels & Apartments (approx. 50 EUR per night)

- Hotel Center: <http://www.hotelcenter.si/HotelCenterWeb/default.aspx>
- B&B Atticus: <http://www.atticus.si/>
- Café & Rooms Maček: <https://sobe-macek.si/en/rezervacija>
- Rooms Ambrožič Ljubljana: <http://rooms-ambrozic.sloveniahotel.net/en/>
- Hotel Meksiko: <http://www.hotel-meksiko.si/>

Mid-range Hotels

- Central Hotel (workshop venue): <http://www.union-hotels.eu/en/central-hotel/>
- Hotel Emonec: <http://www.hotel-emonec.com/>
- Hotel City: <https://www.cityhotel.si/>
- Hotel Park: <http://www.hotelpark.si/#>
- M Hotel (2 km from city centre): <https://www.m-hotel.si/en>

6. Budget

The budget for the Training School was set at 25.000 euro. This budget was meant for 5 trainers and 30 trainees and for accommodation. Fortunately there was no charge for the training facilities, which saved some money for other events:

Travel and accommodation (Trainers)	4003,90 EUR
Grants Trainees	11.734,33 EUR
Local Organizing Support	3120,00 EUR
TOTAL	18.858,23 EUR



ANNEX 1 : Certificate of Participation

CERTIFICATE OF PARTICIPATION

Training School

Tools and methods for creating innovative e-dictionaries

EUROPEAN NETWORK OF e-LEXICOGRAPHY – ENeL

Univesity

Lisbon 6–10 July 2015

Name participant:

Institute:

Country:

Rute Costa

Training School Manager

Simon Krek

Chair of Working Group 3

ANNEX 2: Attendance List