

Research Institute for Linguistics  
Hungarian Academy of Sciences  
CLARA 238405, SP3-People-ITN

# **Report on the CLARA Thematic Training Course on Processing Morphologically Rich Languages**

11-15 April, 2011

Budapest  
11 July, 2011



## 1 Description

The University of Helsinki and the Research Institute for Linguistics, Hungarian Academy of Sciences hosted a one week training course on computational processing of morphologically rich languages from 11-15 April 2011 at RIL in Budapest, Hungary, with the following invited speakers:

- András Kornai, Computer and Automation Research Institute, Hungarian Academy of sciences
- Péter Mihajlik, Budapest University of Technology and Economics, Department of Telecommunication and Media Informatics
- Sjur Moshagen, Norwegian Saami Parliament
- Gábor Prószték, MorphoLogic Ltd.
- Balázs Tarján, Budapest University of Technology and Economics, Department of Telecommunication and Media Informatics
- Veronika Vincze, University of Szeged, Department of Informatics, Human Language Technology Group

**Aim and focus** Morphologically-rich languages like Turkish, Finnish, Hungarian, Sámi etc. present significant challenges for natural language processing applications due to their relatively free word order and highly productive morphological and morphophonological processes (inflection, agglutination, compounding, vowel harmony). The course worked on problems due to dictionary size, sparse data, poor language model probability estimation, high out-of-vocabulary rate and information gaps on related lexical items.

The course introduced advanced modelling techniques addressing these problems, such as decomposition of complex word forms into smaller units, relating inflectional variants to root forms (lemmatization), methods for optimizing the selection of units at different levels of processing, novel probability estimation techniques, and the creation of a new class of data resources and annotation tools. The course concluded with an assessment of present day standard techniques and a demonstration of practical applications focusing primarily on Hungarian and Finnish.

The course focused on practical issues and problem-solving strategies when building large-scale LT tools for morphologically complex languages. Students received hands-on tutorials and exercises to get a better understanding of the complexities involved when building such tools, and also to get a better grasp of the methods taught to handle this complexity.

The course web page is available at the following address:  
<http://www.ling.helsinki.fi/events/CLARA-FSM/>



## 2 Timetable

	Monday	Tuesday	Wednesday	Thursday	Friday
9.00-12.00	<b>S1</b> Sjur Moshagen: Introduction, installation setup, available course LT resource overview	<b>S1</b> Sjur Moshagen: Morphology	<b>S1</b> Sjur Moshagen: Compounds & derivations	<b>S1</b> Gábor Prószéky: Problems and solutions in computational morphology applications of highly inflectional languages	<b>S1</b> András Kornai: Finite state methods in lexical semantics
12.00-14.00	lunch break				
14.00-17.00	<b>S2</b> Sjur Moshagen: Working with a full scale lexicon	<b>S2</b> Sjur Moshagen: Mor- phophonology	<b>S2</b> Sjur Moshagen: Practical applications	<b>S2</b> Veronika Vincze: Issues related to syntax- morphology interface	<b>S2</b> Péter Mihajlik and Balázs Tarján: Large vocabulary continuous speech recognition in agglutinative languages



### 3 List of participants

Name	Affiliation	Email
Zeeshan Ahmed	University College Dublin	zeeshan.ahmed@ucdconnect.ie
Gabor Csernyi	University of Debrecen	gabor.csernyi@arts.unideb.hu
Gregoire Detrez	Göteborg University	detrez@chalmers.se
Bamba Dione	University Of Bergen	dione.bamba@iie.uib.no
Senka Drobac	University of Helsinki	senka.drobac@helsinki.fi
Nathan Green	Charles University	clara@nathangreen.com
Mark Kane	University College Dublin	mark.kane@ucdconnect.ie
Márton Károly	University of Pécs	harczymarczy@gmail.com
Jianqiang Ma	University of Tübingen	jma@sfs.uni-tuebingen.de
Udochukwu Ogbureke	University College Dublin	udo.kalu@yahoo.es
Tommi Pirinen	University of Helsinki	tommi.pirinen@helsinki.fi
Loganathan Ramasamy	Charles University in Prague	ramasamy@ufal.mff.cuni.cz
Teemu Ruokolainen	Aalto University, Department of Information and Computer Science	teemu.ruokolainen@tkk.fi
Miikka Silfverberg	University of Helsinki	miikka.silfverberg@helsinki.fi
Éva Székely	University College Dublin	Eva.Szekely@ucdconnect.ie
Amalia Zahra	University College Dublin	amalia.zahra@ucdconnect.ie



## 4 Workshop organization

Organization and administration were provided by the hosting institute. The course took place in the main lecture hall at the Research Institute for Linguistics with on site registration. Free wireless internet was available in the institute, as well as guest login to a central server with all necessary software pre-installed, with account details provided upon registration.

For those participants and lecturers who required lunch (Table 1) was offered at a nearby self-service restaurant. Lunch and the workshop dinner (held on Wednesday, April 13) were covered from the course budget.

	Monday	Tuesday	Wednesday	Thursday	Friday	Total	Refund
András Kornai					2465	2465	2465
Péter Mihajlik	1500				1700	3200	3200
Sjur Moshagen	1160	1180		1890	1380	5610	5610
Balázs Tarján	1420				1440	2860	2860
Veronika Vincze				1270	1280	2550	2550
Zeeshan Ahmed	1050	1190	1080	1260	1700	6280	6280
Gabor Csernyi	1370	1305	1370		1720	5765	5765
Gregoire Detrez		1320		1620	3220	6160	6160
Bamba Dione	1350	1645	1870	1700	1480	8045	8045
Nathan Green	1800	1530	1560			4890	4890
Mark Kane	1000	1100	1280	830	1700	5910	5910
Márton Károly	750	630	1270	1390	1010	5050	5050
Jianqiang Ma	1270	1900	1810	1890	1710	8580	8580
Udochukwu Ogbureke	1440	1870	1390	1640	1905	8245	8245
Loganathan Ramasamy	1430	1710	1195	1335	1555	7225	7225
Amalia Zahra	1060	1100	980	1190	1065	5395	5395
Total:	16600	16480	13805	16015	25330	88230	88230

Table 1: Detailed lunch costs for each participant/lecturer in Hungarian forints.



## 5 Additional documents

A scanned copy of the registration sheet for participants (Figure 1) and lecturers (Figure 2), and a separate record on attendance by all CLARA external participants (Figure 3) and lecturers (Figure 4) are attached to this report.



### Registration

Name	Affiliation	Dinner	Signature
Zeeshan Ahmed	University College Dublin	✓	<i>Zeeshan</i>
Gabor Csernyi	University of Debrecen	✓	<i>G. Csernyi</i>
Gregoire Detrez	Gothenburg University	✓	<i>Gregoire</i>
Bamba Dione	University Of Bergen	✓	<i>Bamba Dione</i>
Senka Drobac	University of Helsinki	✓	<i>Senka Drobac</i>
Nathan Green	Charles University	✓	<i>Nathan Green</i>
Mark Kane	University College Dublin	✓	<i>Mark Kane</i>
Oleg Kapanadze	<del>University of Saarland</del>		
Márton Károly	<i>University of Pécs</i>	✓	<i>M. Károly</i>
Zoltán Ludány	<del>University of Pécs</del>		
Jianqiang Ma	University of Tübingen	✓	<i>Jianqiang Ma</i>
Udochukwu Ogbureke	University College Dublin	✓	<i>Udochukwu Ogbureke</i>
Tommi Pirinen	University of Helsinki	✓	<i>Tommi Pirinen</i>
Loganathan Ramasamy	Charles University in Prague	✓	<i>Loganathan Ramasamy</i>
Teemu Ruokolainen	Aalto University, Department of Information and Computer Science	✓	<i>Teemu Ruokolainen</i>
Miikka Silfverberg	University of Helsinki	✓	<i>M. Silfverberg</i>
Éva Székely	University College Dublin	✓	<i>Éva Székely</i>
Shafqat Virk	<del>Graduate School of Language Technology, University of Gothenburg</del>		
Amalia Zahra	University College Dublin	✓	<i>Amalia Zahra</i>

CLARA Thematic Training Course on Processing Morphologically Rich Languages

Figure 1: Registration for participants.





### Lecturers


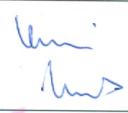



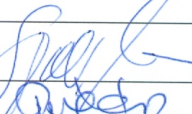

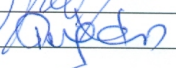




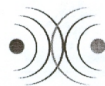
Name	Affiliation	Dinner	Signature
András Kornai	Computer and Automation Research Institute, Hungarian Academy of sciences		
Péter Mihajlik	Budapest University of Technology and Economics Department of Telecommunication and Media Informatics		
Sjur Moshagen	Norwegian Saami Parliament		
Gábor Prószéky	MorphoLogic		
Balázs Tarján	Budapest University of Technology and Economics Department of Telecommunication and Media Informatics		
Veronika Vincze	University of Szeged, Department of Informatics, Human Language Technology Group		

Figure 2: Registration for lecturers.

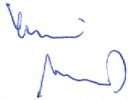
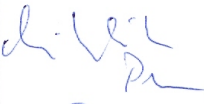
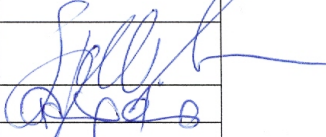
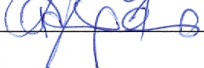
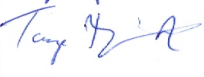





Participants			
Name	Affiliation	Attendance (num. of days)	Signature
Zeeshan Ahmed	University College Dublin	5	<i>Zeeshan Ahmed</i>
Gabor Csernyi	University of Debrecen	5	<i>Gabor Csernyi</i>
Gregoire Detrez	Göteborg University	5	<i>Gregoire Detrez</i>
Mark Kane	University College Dublin	5	<i>Mark Kane</i>
Márton Károly	University of Pécs	5	<i>Márton Károly</i>
<del>Zoltán Ludány</del>	<del>University of Pécs</del>		
Udochukwu Ogbureke	University College Dublin	5	<i>Udochukwu Ogbureke</i>
Teemu Ruokolainen	Aalto University, Department of Information and Computer Science	5	<i>Teemu Ruokolainen</i>
Éva Székely	University College Dublin	4	<i>Éva Székely</i>
Amalia Zahra	University College Dublin	5	<i>Amalia Zahra</i>

Figure 3: Participants outside CLARA.



Lecturers			
Name	Affiliation	Attendance (num. of days)	Signature
András Kornai	Computer and Automation Research Institute, Hungarian Academy of sciences	1	
Péter Mihajlik	Budapest University of Technology and Economics Department of Telecommunication and Media Informatics	4	
Sjur Moshagen	Norwegian Saami Parliament	4+1	
Gábor Prószéky	MorphoLogic	1	
Balázs Tarján	Budapest University of Technology and Economics Department of Telecommunication and Media Informatics	3	
Veronika Vincze	University of Szeged, Department of Informatics, Human Language Technology Group	2	

CLARA Thematic Training Course on Processing Morphologically Rich Languages  
11-15 April, 2011

Figure 4: Lecturers outside CLARA.