ELEXIS: Technical and social infrastructure for lexicography

Anna Woldrich, Teja Goli, Iztok Kosem, Ondřej Matuška and Tanja Wissik

Since the European lexicographic community was brought together by the European Network of e-Lexicography (ENeL) COST action (http://elexicography.eu/) in 2013–2017, the following needs have become apparent: the flow of broader and more systematic exchange of expertise; the establishment of common standards and solutions; the development and integration of lexicographic resources; and the wide-scale application of these quality resources to wider research communities. This has resulted in launching the four-year H2020 infrastructure project ELEXIS, European Lexicographic Infrastructure in February 2018 (extended for six months until July 2022).

ELEXIS brings together research and industrial partners from various fields, such as the Semantic Web, Artificial Intelligence, Natural Language Processing (NLP) and Digital Humanities, thus supporting developments in (e-)lexicography in order to open up dictionary data and enable access to lexicographic standards, methods, data and tools.

Among the most obvious outputs of the project are the tools and services it offers. In its first two years, ELEXIS has been enriched by seven different tools which were either developed as part of the project or made freely accessible through its infrastructure, and by the end of the project, the ELEXIS infrastructure is planned to enable and support the whole dictionary creation process. The tools and services already available include:

Sketch Engine. This corpus query system, which existed prior to the project, was one of the first tools made freely accessible to academics and observer institutions in ELEXIS. It includes over 500 preloaded corpora and analysis functions, such as concordancing, building wordlists, compiling word sketches, thesauri and automatic dictionary drafting. https://sketchengine.eu/elexis/

Lexonomy. Another infrastructure component which already existed before ELEXIS but whose further comprehensive development continues within the project. This is a cloud-based dictionary-writing and online-publishing system that interacts closely with Sketch Engine. For example, Sketch Engine can push lexicographic data into Lexonomy to create automatically generated dictionary drafts and Lexonomy can pull data from Sketch Engine's corpora during the entry editing process. https://lexonomy.eu/

Anna Woldrich is communication expert at the Austrian Academy of Sciences (ACDH-CH) and has worked previously as social editor and campaign manager. She graduated at Universität Wien/Universitá degli Studi di Siena in mass media and communication studies, focusing on radio, broadcasting, marketing, communication research and communication theory. As a part of ELEXIS project, she is responsible for planning, managing and monitoring on- and offline communication activities, and manages the social media channels and content-management on the ELEXIS website.
anna.woldrich@oeaw.ac.at

Elexifier. A brand new cloud-based dictionary conversion service, using advanced XML parsing and machine learning techniques to help convert PDF and XML dictionary data into a standardized machine-readable format. Users can upload PDF and custom XML dictionaries, define mapping rules for XML transformation or create a machine learning training set for PDF conversion and download the transformed XML or PDF dictionary in a TEI-compliant file format based on the Elexis Data Model[1]. https://elexifier.elex.is/

VerbAtlas. A novel large-scale manually-crafted semantic resource for wide-coverage, intelligible and scalable semantic role labeling. The goal of VerbAtlas is to manually cluster WordNet synsets that share similar meanings into sets of semantically-coherent frames, available both for download and via a RESTful API, featuring resources such as PropBank and BabelNet. http://verbatlas.org/

SyntagNet. A manually-curated large-scale lexical-semantic combination database which associates pairs of concepts with pairs of co-occurring words. The goal of SyntagNet is to capture sense distinctions evoked by syntagmatic relations, hence providing information which complements the essentially paradigmatic knowledge shared by currently available lexical knowledge settings such as WordNet. http://syntagnet.org/

NAISC. A tool for linking datasets. NAISC takes as input 2 RDF documents (referred to as 'left' and 'right) and outputs an alignment (set of RDF triples) between these two documents. It typically relies on a configuration, which is a JSON document. https://youtube.com/watch?v=maYEv8rG0_k

Elexifinder. A search tool dedicated to helping lexicographers and researchers find scientific output in lexicography and related fields. Elexifinder enables users to search through papers and videos, using concepts, that is words or sets of words with a Wikipedia page, and various other conditions, for example source (conference, etc.), author, language, etc. Each paper/video is linked to its page where the user can download or view it. https://elex.is/tools-and-services/elexifinder/; http://er.elex.is/.

Lexicographic news feed. A service using the Event Registry API to extract recent news articles related to lexicography. Articles are extracted from 30,000 news sources, supporting over. 35 languages. https://elex.is/tools-and-services/lexicographic-news/.

1 Information on the ELEXIS Data Model is available in the recordings of the ELEXIS Observer Event 2019: http://videolectures.net/elexisobserver2019_tiberius_data_model/

Teja Goli is an assistant at the Artificial Intelligence Laboratory at Jožef Stefan Institute and at the University of Ljubljana, where she has finished her master's degree in Translation at the Faculty of Arts. Her research interests include translation, corpus linguistics and lexicography. In the ELEXIS project, she is mainly responsible for contact with observers and managing website content.
teja.goli@ijs.si

Ondřej Matuška oversees sales and marketing activities and external communication at Lexical Computing, and is the main point of contact for information about and user support for Sketch Engine.
ondrej.matuska@sketchengine.co.uk

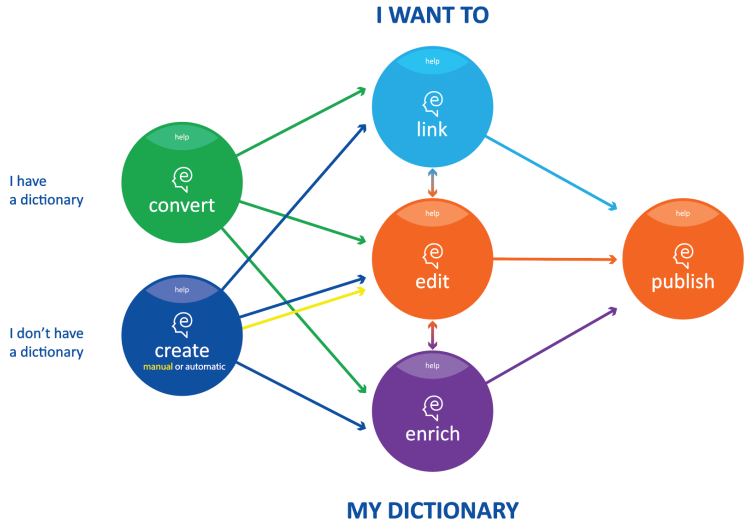

Image 1. ELEXIS offers a user-friendly way to create dictionaries or edit and publish existing ones

More tools and services as well as instruction manuals will be added during the lifetime of the project, accumulating to a full-scale service for a user-friendly dictionary publication process (cf. Image 1).

Besides this extensive technical infrastructure, ELEXIS provides a social infrastructure to foster cooperation and support knowledge exchange among lexicographic communities. Additionally, it is bridging the gap between lesser-resourced languages and those with higher e-lexicographic expertise. One aspect of this social infrastructure is organizing training sessions and workshops at conferences as well as summer/spring schools all over Europe (https://elex.is/all-events/). Due to COVID-19 restrictions, several events had to be canceled this year, but we have managed to overcome the obstacle prohibiting face-to-face interaction for community building by moving several activities online. As part of the GlobaLex 2020 Workshop on Linked Lexicography at the Language Resource and Evaluation Conference (LREC 2020), ELEXIS organized the first shared task on monolingual word-sense alignment (MWSA). While the workshop itself had to be cancelled, the papers and the results are available as part of the proceedings (https://aclweb.org/anthology/volumes/2020.globalex-1/). The goal was to find senses in two monolingual dictionaries (in the same language), that describe the same concept. The MWSA task made use of data in 15 languages from ELEXIS partners and observers. The participants developed strong systems with the overall best system scoring 84% accuracy in sense alignment. https://elex.is/mwsa2020/.

Furthermore, ELEXIS supports individual researchers and research teams via trans-national access, enabling them to reach facilities

Iztok Kosem (PhD) is Research Associate at Jožef Stefan Institute and at the University of Ljubljana. His main areas of research are lexicography and lexicogrammar, corpus linguistics, crowdsourcing, and computer-aided language learning and teaching. In ELEXIS, he has the role of Community Manager, and he is heavily involved in the development of Elexifinder, Lexonomy and games with a purpose (gamification).
iztok.kosem@ijs.si

and lexicographic resources which are not fully or easily accessible online or where professional on-site expertise is needed. Researchers, scholars and students are invited to apply for a fully-funded short- or long-term research visit to leading lexicographic institution partners (https://elex.is/grants-for-research-visits/). Calls for visiting grants are launched twice a year, in summer and in winter, amounting to seven calls in total during the project period. The travel grant reports as well as mini-interviews with the respective winners from various countries all over Europe are available at https://elex.is/travel-grant-reports/.

While individual researchers can participate through travel grants, institutions are invited to join the ELEXIS network via observer status (https://elex.is/join-as-observer). Observing institutions may request new customized lexicographic data or have their existing data enriched and expanded with both monolingual and multilingual information, Moreover, they can access the ELEXIS cloud, tools and open-access resources as well as resources in the partner and observer's area of the cloud. Observers are notified about newly developed tools, services and activities (e.g. hackathons, tool demo sessions, etc.) aimed at improving and enriching their own lexicographic data. To keep up

The first Lexonomy Hackaton took place in Brno on 23-25 April 2019

a sustainable infrastructure after the end of the project in 2022, the observer status guarantees the possibility to participate actively in the post-project stage.

To this end, ELEXIS organized an Observer Event in early 2019, dedicated to inform representatives of various lexicographic institutions on its activities (https://elex.is/observer-event/).

Tanja Wissik is a senior scientist and project leader at the Austrian Centre for Digital Humanities and Cultural Heritage of the Austrian Academy of Sciences and she teaches at the University of Graz and University of Vienna. She holds a PhD in Translation Studies from the University of Vienna and in the last couple of years she has been working in a number of European and national research projects in the field of language resources, text technologies and DH methods.
tanja.wissik@oeaw.ac.at

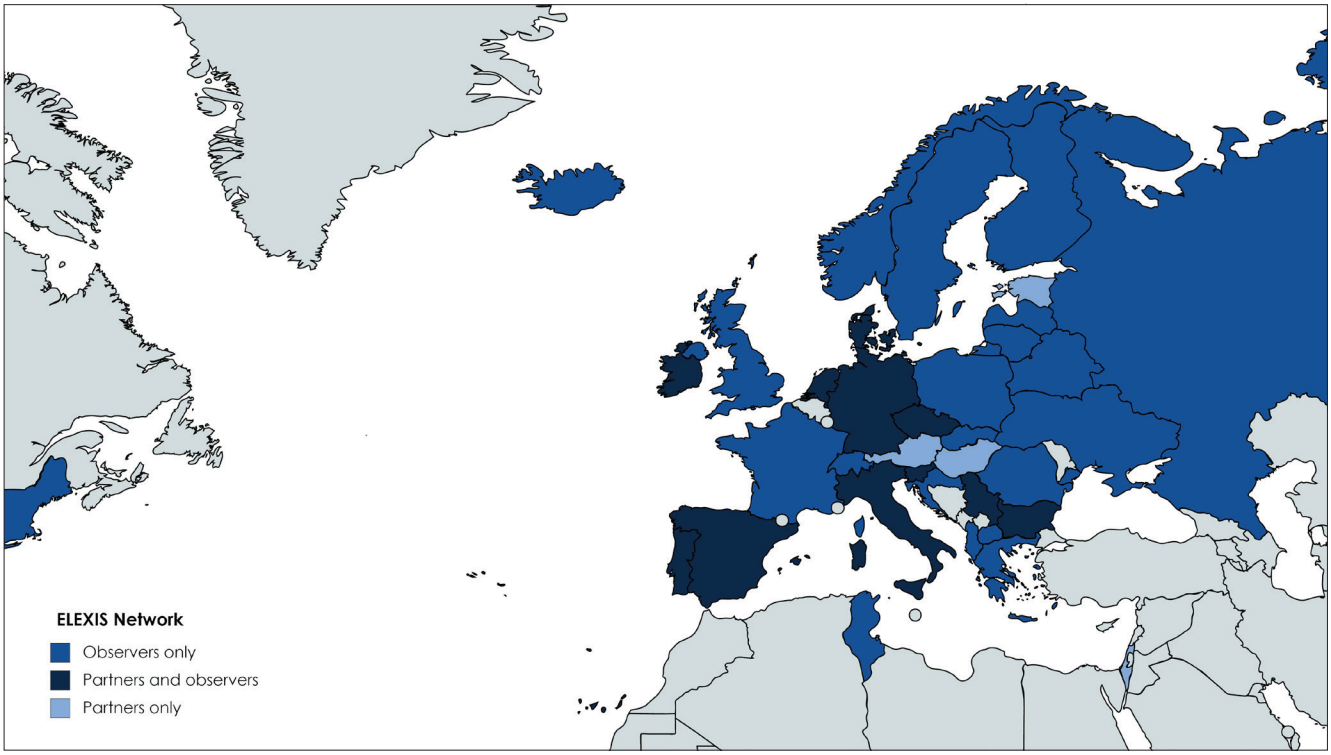

Image 2. Overview of the ELEXIS network in June 2020

Institutions from all over Europe (and beyond) have been joining the network: as of June 2020, the ELEXIS community is made up of 17 partner and 50 observer institutions from 35 different countries (cf. Image 2; https://elex.is/observers/). In addition, ELEXIS is running a campaign on social media, describing the characteristics of each observing institution – all portraits are collected in the #elexisobserver moment on Twitter.

Since community building is a key factor for ELEXIS, it is important to assess the experience and opinions regarding the project's intermediate outcomes. This is a way to reflect on the work done so far as well as to fine-tune the final outcomes to respond best to the needs of the community. Thus, the ELEXIS impact survey was launched in May 2020, containing 16 questions on different aspects of the technical and social infrastructures. The results have shown that 79% (n=123) of the respondents already knew ELEXIS or were following its activities actively. For most respondents the most important aspects of ELEXIS are the tools and services as well as open access and open data, followed by training and education, knowledge exchange and community building (cf. Image 3).

elexis
european lexicographic infrastructure

Horizon 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731015.

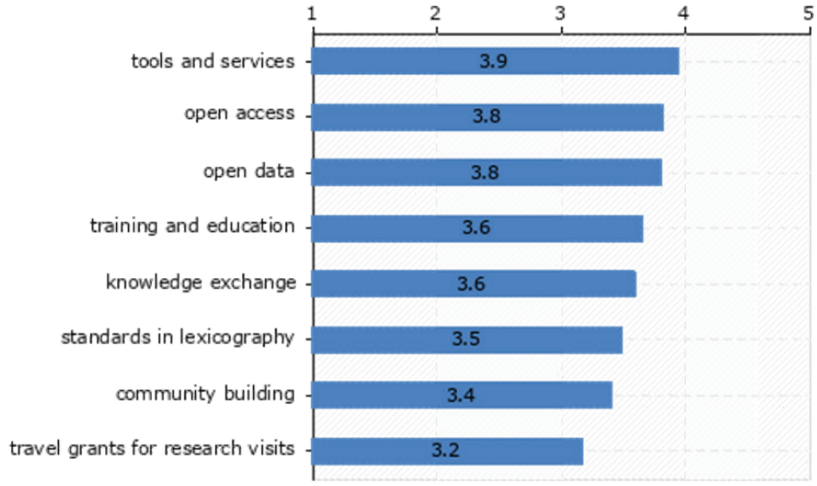

Image 3. Usefulness of ELEXIS services for those who are familiar with the network (Q12, N=97)

Although some respondents did not know ELEXIS before, we were interested to find out how useful specific aspects of the infrastructure might be to them. These turned out to include access to the corpus query tool Sketch Engine, open data and open access, as well as knowledge exchange, training and education (cf. Image 4).

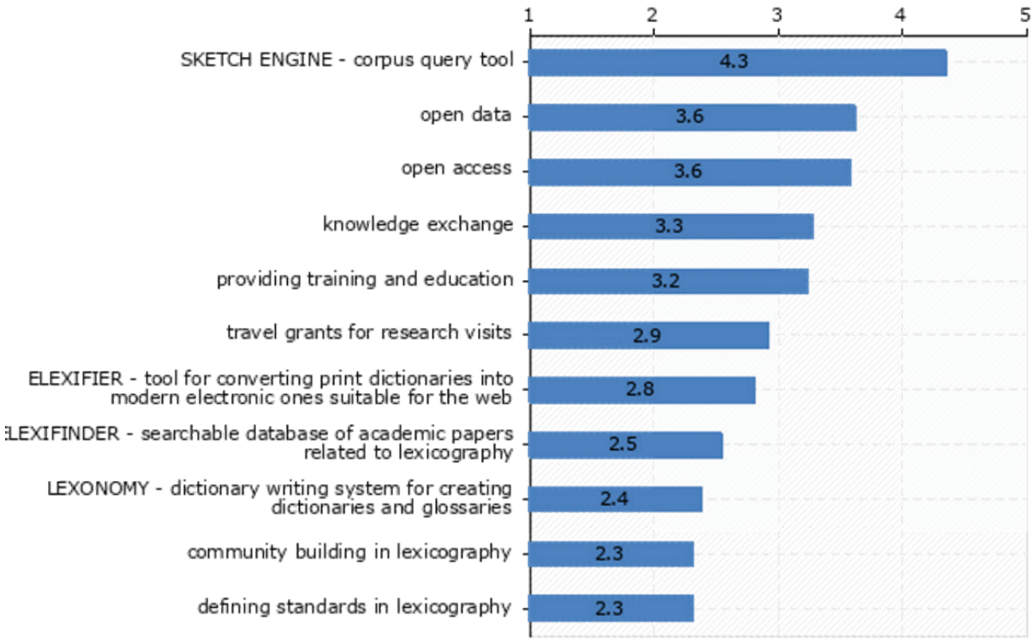

Image 4. Potential usefulness of ELEXIS services for those who don't know the network (Q6, N=26)

The full survey as well as other project reports are available at https://elex.is/deliverables/.

Additionally, all conference papers, peer-reviewed articles and journal articles published in the course of the project with ELEXIS are available on Zenodo.

Partners

Original name	English name	Country
Österreichische Akademie der Wissenschaften	The Austrian Academy of Sciences	Austria
ИНСТИТУТ ЗА БЪЛГАРСКИ ЕЗИК	Institute for Bulgarian Language Prof Lyubomir Andreychin	Bulgaria
Lexical Computing CZ sro	Lexical Computing CZ sro	Czech Republic
Det Danske Sprog- og Litteraturselskab	The Society for Danish Language and Literature	Denmark
Center for Sprogteknologi (CST) Institut for Nordiske Studier og Sprogvidenskab	The Centre for Language Technology at the Department of Nordic Research, University of Copenhagen	Denmark
Eesti Keele Instituut	Institute of the Estonian Language	Estonia
Universität Trier	The Trier Center for Digital Humanities	Germany
Magyar Tudományos Akadémia Nyelvtudományi Intézetének	Research Institute for Linguistics at the Hungarian Academy of Sciences	Hungary
The National University of Ireland, Galway/ OÉ Gaillimh	The National University of Ireland, Galway	Ireland
ק מילונים בע"מ	K Dictionaries Ltd	Israel
Sapienza Università di Roma	The Sapienza University of Rome	Italy
Consiglio Nazionale delle Ricerche – Istituto di Linguistica Computazionale "A Zampolli"	The Institute for Computational Linguistics "A Zampolli"	Italy
Universidade Nova de Lisboa – Faculdade de Ciências Sociais e Humanas	Universidade NOVA de Lisboa – The NOVA School of Social Sciences and Humanitie	Portugal
ентар за дигиталне хуманистичке науке	The Belgrade Center for Digital Humanities	Serbia
Inštitut "Jožef Stefan"	"Jožef Stefan" Institute	Slovenia
Real Acamedia Espanola	The Royal Spanish Academy	Spain
Instituut voor de Nederlandse Taal	Dutch Language Institute	The Netherlands

ELEXIS on GitHub

https://github.com/elexis-eu

Lexonomy
https://github.com/elexis-eu/lexonomy
Elexifinder
https://github.com/elexis-eu/elexifinder
elexifier-api
https://github.com/elexis-eu/elexifier-api
elexifier
https://github.com/elexis-eu/elexifier
dictionary service
https://github.com/elexis-eu/dictionary-service
MWSA
https://github.com/elexis-eu/MWSA
NAISC
https://github.com/insight-centre/naisc
word games
https://github.com/elexis-eu/word-games
elexis-rest
https://github.com/elexis-eu/elexis-rest
elexifier-pdf
https://github.com/elexis-eu/elexifier-pdf
tei2ontolex
https://github.com/elexis-eu/tei2ontolex
CrossTheWord
https://github.com/elexis-eu/CrossTheWord
ocd
https://github.com/elexis-eu/ocd
D3.1
https://github.com/elexis-eu/D3.1

Observers (June 2020)

	Original name / English name	Country
1	Institut za hrvatski jezik i jezikoslovlje / Institute for the Croatian Language and Linguistics	Croatia
2	Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE) / Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE)	Italy
3	Univerzitet u Beogradu, Rudarsko-geološki fakultet / Univesity of Belgrade, Faculty of Mining and Geology	Serbia
4	SIL International / SIL International	International (US)
5	Лексикографски центар при Македонската академија на науките и уметностите / Lexicographic Centre at the Macedonian Academy of Sciences and Arts	North Macedonia
6	Kotimaisten kielten keskus / Institute for the Languages of Finland	Finland
7	Київський університет імені Бориса Грінченка / Borys Grinchenko Kyiv University	Ukraine
8	Институт по информационни и комуникационни технологии към Българската академия на науките / Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences (IICT-BAS)	Bulgaria
9	Институт лингвистических исследований Российской академии наук / Institute for Linguistic Studies, Russian Academy of Sciences	Russia
10	Universitatea de Vest din Timisoara / West University of Timisoara	Romania
11	Universitetsbiblioteket ved Universitetet i Bergen / University of Bergen Library	Norway
12	Sveučilište Jurja Dobrile u Puli / Juraj Dobrila University of Pula	Croatia
13	Filozofski fakultet, Sveučilište u Rijeci / Faculty of Humanities and Social Sciences, University of Rijeka	Croatia
14	Institut de Recherche et d'Histoire des Textes (CNRS) / Institut de Recherche et d'Histoire des Textes (CNRS)	France
15	Dipartimento di filologia, letteratura, linguistica / Department of Philology, Literature and Linguistics	Italy
16	Vytauto Didžiojo universitetas, Kompiuterinės lingvistikos centras / Vytautas Magnus University, Centre of Computational Linguistics	Lithuania
17	Uniwersytet im. Adama Mickiewicza w Poznaniu / Adam Mickiewicz University in Poznań	Poland
18	Institutul de Filologie Română „A. Philippide" / "A. Philippide" Institute of Romanian Philology	Romania
19	Lietuvių kalbos institutas / The Institute of the Lithuanian Language	Lithuania
20	ZRC SAZU Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti / ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	Slovenia
21	Stofnun Árna Magnússonar í íslenskum fræðum / The Árni Magnússon Institute for Icelandic Studies	Iceland
22	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, v. v. i. / Ľudovit Štúr Institute of Linguistics of the Slovak Academy Sciences	Slovakia
23	Sveučilište u Zagrebu, Filozofski fakultet / University of Zagreb, Faculty of Humanities and Social Sciences	Croatia
24	Schweizerisches Idiotikon / Swiss Idiotikon	Switzerland

	Original name / English name	Country
25	Universidad de Castilla-La Mancha / University of Castilla–La Mancha	Spain
26	Institute for Applied Linguistics, Eurac Research / Institute for Applied Linguistics, Eurac Research	Italy
27	UPV/EHU University of the Basque Country / UPV/EHU University of the Basque Country	Spain
28	Instytut Neofilologii – Państwowa Wyższa Szkoła Zawodowa w Raciborzu / Institute of Modern Language Studies – State University of Applied Sciences in Racibórz	Poland
29	Institut Superior d'Investigació Cooperativa – IVITRA (Universitat d'Alacant) / Higher Institute Of Cooperative Research – IVITRA (University of Alicante)	Spain
30	Foras na Gaeilge / Foras na Gaeilge	Ireland
31	Stiechting Limbörgse Academie / Limburgish Academy Foundation	The Netherlands
32	Zavod za lingvistička istraživanja Hrvatske akademije znanosti i umjetnosti / Linguistics Research Institute of the Croatian Academy of Sciences	Croatia
33	Latvijas Universitātes Matemātikas un informātikas institūts / Institute of Mathematics and Computer Science, University of Latvia	Latvia
34	Center za jezikovne vire in tehnologije, Univerza v Ljubljani / Centre for Language Resources and Technologies, University of Ljubljana	Slovenia
35	Academia das Ciências de Lisboa / Lisbon Academy of Sciences	Portugal
36	Canolfan Uwchefrydiau Cymreig a Cheltaidd Prifysgol Cymru / University of Wales Centre for Advanced Welsh & Celtic Studies	United Kingdom
37	Institut für Deutsche Sprache / Institute for the German Language	Germany
38	Svenska Akademien / Swedish Academy	Sweden
39	Cologne Center for Humanities / Cologne Center for Humanities	Germany
40	Ústav Českého národního korpusu / Institute of the Czech National Corpus	Czech Republic
41	Mykolo Romerio Universitetas / Mykolas Romeris University	Lithuania
42	Institute for Language and Speech Processing, ATHENA R.C. / Ινστιτούτο Επεξεργασίας του Λόγου, Ε.Κ. ΑΘΗΝΑ	Greece
43	Universitatea de Medicină, Farmacie, Științe și Tehnologie „George Emil Palade" din Târgu Mureș / George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures	Romania
44	Гродзенскі дзяржаўны ўніверсітэт імя Янкі Купалы / Гродненский государственный университет имени Янки Купалы / Yanka Kupala State University of Grodno	Belarus
45	Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) / Berlin-Brandenburg Academy of Sciences and Humanities	Germany
46	Univerzitet u Kragujevcu, Filološko-umetnički fakultet / University of Kragujevac, Faculty of Philology and Arts	Serbia
47	Fakulteti i Historisë dhe i Filologjisë / Faculty of History and Philology	Albania
48	Instituti i Gjuhësisë dhe i Letërsisë / Institute of Linguistics and Literature	Albania
49	Universidad de Alcala / University of Alcala	Spain
50	Dansk Sprognævn / The Danish Language Council	Denmark