LISBON SUMMER SCHOOL IN LINGUISTICS

PhD Programme

Linguistics – Knowledge, Representation and Use

Lisboa, 5-9 July 2021

Venue
Faculdade de Ciências Sociais e Humanas – NOVA
Avenida de Berna, 26-C
1069-061 Lisboa – Portugal

AREA 2: TERMINOLOGY AND LEXICOGRAPHY

Course 1: **The past, present and future, of African language dictionaries - from Eurocentric to Afrocentric compilations** – D.J. Prinsloo – Faculty of Humanities, University of Pretoria, South Africa (classes by videoconference, ENGLISH)

COURSE DESCRIPTION

The aim of the seminar is to give a perspective on the past, present and future of African language lexicography. It will be indicated that African lexicography is in a developing stage and mostly paper based. Therefore, the question is whether African lexicography should make a fresh start on electronic dictionary compilation or whether the ideal of paper dictionaries of high lexicographic achievement should be reached first. Attention will also be given to the movement from Eurocentric to Afrocentric lexicography where dictionaries were originally compiled by Europeans but today increasingly by Africans — dictionaries for African languages by Africans in Africa. The future of dictionaries and in particular the potential threat of internet data to the future of dictionaries will be outlined.

Selected references

Prinsloo, D.J. (2019). Lexicography: A perspective on the past, present and future of lexicography with specific reference to Africa. Keynote presentation Conference *Proceedings of AsiaLex 2019, 13th International Conference of the Asian Association for Lexicography* 19 to 21 June 2019. Istanbul University Congress and Culture Center. Istanbul Turkey, pp. 148-160

Prinsloo, D.J. / Zondi, Nompumele (2020). From postcolonial African language lexicography to globally competitive e-lexicography in Africa. In: Russell H Kaschula & H Ekkehard Wolff (eds) *African Languages in Knowledge Societies. The transformative power of languages*. Cambridge University Press, pp. 259 - 274

Prinsloo, D.J. (2019). Detection and lexicographic treatment of salient features in e-dictionaries for African languages. *International Journal of lexicography*. 33(3) 269-287. Advanced publication: ecz031, https://doi.org/10.1093/ijl/ecz031.

Prinsloo, D.J. / Prinsloo, J.V. / Prinsloo, Daniel (2017). African Lexicography in the Era of the Internet. Pedro A. Fuertes Olivera (Ed.). *The Routledge Handbook of Lexicography*. London: Routledge, pp. 487-502.

Prinsloo, D.J. / Taljard, Elsabé. (2017). Afrikataalleksikografie: gister, vandag en môre [African Language Lexicography: yesterday, today and tomorrow]. *Lexikos 2017*. (427-456). IF .667

Biography

Prof. Daniel Jacobus Prinsloo is an internationally renowned professor and researcher in African Languages at the University of Pretoria. With more than 280 research outputs on African Language Lexicography, Linguistics and Human Language technology, Prof. Prinsloo received the Outstanding Academic Performer Award in 2012-2014, 2009-2011, 2006-2008, 2003-2005 and 2000-2002, the Pan South African Language Board award in 2010, among others. Prof. Prinsloo is a founder member of the African Association for Lexicography and was Board member of the Sepedi National Lexicography Unit and of ALASA, besides having participated as editor and co-founder of the prestigious journal Lexikos. Currently teaching at the Department of African Languages at the Faculty of Humanities of University of Pretoria, Prof. Prinsloo has been collaborating with many universities and research centers in South Africa and around the world, namely the Centre for Text Technology, NWU, the South African Centre for Digital Language Resources, Ghent University, Belgium, Guangzhou University of Foreign Studies, China, Institut fűr Maschinelle Sprachverarbeitung, University of Stuttgart.

Course 2: **Linguistic Digital Data in Terminology and Lexicography** – Raquel Amaro / Rute Costa – CLUNL, NOVA FCSH (classes on site, PORTUGUESE/ENGLISH).

COURSE DESCRIPTION

Terminological and lexicographical work aim to compile and systematize linguistic and conceptual information and make it available to end users – which can be both humans and machines – ideally in a format that can be used, reused, and shared. In this context, standards and formalization play a major role. Both subject fields have become central disciplines in numerous areas of knowledge, ranging from eHealth to Digital Humanities, passing through language teaching and base resources development for Computational Linguistics, Natural Language Processing, and Artificial Intelligence. However, although Terminology and Lexicology/Lexicography deal with lexical units – words and terms –, they take up different perspectives in what concerns theories, methods, and end results. The main goals of the seminar are, thus, the confrontation of approaches and the opening to the diversity of work and research developed in the areas of Terminology, Lexicography and Computational Lexical Semantics, and the demonstration of how linguistic digital data are conceived, treated and used in these fields

Selected references

Amaro, Raquel (2018). Integrating Prepositions in WordNets: Relations, Glosses and Visual Description. In: I. Kernerman, S. Krek, eds., *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*. Miyazaki: European Language Resources Association (ELRA), pp. 66-74. ISBN 979-10-95546-28-3.

Amaro, Raquel / Mendes, Sara (2016). Lexicologia e linguística computacional. In: A. M. Martins, E. Carrilho, eds., Manual de Linguística Portuguesa. Berlim: De Gruyter, pp.178-199. ISBN 978-3-11-037448-3.

Costa, Rute / Carvalho, Sara / Salgado, Ana / Simões, Alberto / Tasovac, Toma (2020). "Ontologie des marques de domaines appliquée aux dictionnaires de langue générale", in [éditeur : Xavier Blanco] La lexicographie en tant que méthodologie de recherche en linguistique Revue de Philologie Française et Romane - Langue(s) & Parole, n. 5. Mons: Edition du CIPA. pp. 201-230. ISSN papier 2466-7757, ISSN numérique 2684-6691.

Gantar, Polona / Colman, Lut / Parra Escartín, Carla / Martínez Alonso, Hector (2019) "Multiword Expressions: Between Lexicography and NLP". In *International Journal of Lexicography*, Volume 32, Issue 2, June 2019, pp. 138–162, https://doi.org/10.1093/ijl/ecy012

Roche, Christophe / Costa, Rute / Carvalho, Sara / Almeida, Bruno (2019). "Knowledge-based terminological e-dictionaries: The EndoTerm and al-Andalus Pottery projects". *Terminology and e-dictionaries*, Special Issue of Terminology 25:2 (2019). (Editors: Alcina Amparo, Rute Costa, Christophe Roche) Amsterdam / Philadelphia: John Benjamins, pp. 259-291. ISSN 0929-9971|e-ISSN 1569-9994.

Biography

Raquel Amaro: https://www.cienciavitae.pt/portal/pt/1515-3C52-D88C

Rute Costa: https://sites.google.com/fcsh.unl.pt/terminology-rute-costa

---

***Course 3*: Introduction to Linked Open Data in Linguistics –** Thierry Declerck – German Research Center for Artificial Intelligence (DFKI) | Austrian Centre for Digital Humanities at the Austrian Academy of Sciences / Julia Bosque-Gil, University of Zaragoza, Spain (classes on site, if allowed by sanitary conditions, ENGLISH).

COURSE DESCRIPTION

Publishing language resources under open licenses and linking them together has been an area of increasing interest in computer science and in digital humanities areas including linguistics, digital heritage, and e-lexicography. This topic has been widely discussed, presented, and deployed in many workshops, datathons, European projects and at ESSLI summer schools. A particularly strong component of this has been the work conducted within the W3C Ontology-Lexica Community Group, whose OntoLex-Lemon model is a de facto standard for lexical data on the Web. Linked data is an important step towards making linguistic data: i) easily and uniformly queryable, ii) interoperable and iii) sharable over the Web using open standards such as the HTTP protocol and the RDF data model. Thus, this course will give participants both theoretical and practical hands-on experiences with the development of these technologies, to help them to make their data re-usable and more sustainable.

This course has the main goal of giving people in the fields of digital humanities and computational linguistics the theoretical underpinnings as well as practical skills in the topics of linked data and semantic technologies as applied to linguistics and lexical data. After developing a short initial ontology, participants will learn step by step how to represent multilingual data with their ontology and how to ground it linguistically. We will introduce a variety of state-of-the-art multilingual representation formats and application scenarios in which to leverage and exploit multilingual semantic data. Finally, we will detail the connection of lexical and corpus resources. At the end of the class, participants will be able to interact with the Linguistic Linked Open Data (LLOD) cloud for the semantic representation of linguistic data.

OUTLINE

Day 1*: Introduction to Semantic Technologies and Ontology Modelling* – Introduction to ontologies including the Resource Description Framework (RDF), the Resource Description Framework Schema (RDFs) and the Web Ontology Language (OWL)

Day 2*:* Details of the Lexicon Model for Ontologies (*lemon)* approach

- Presentation of how to linguistically ground semantic technologies with *lemon*. Inspection of the five modules of *lemon*, with a focus on its core module: OntoLex *(Ontology-Lexicon Interface)*

Day 3*:* Presentation of lexical representation formats related to OntoLex-Lemon, such as

LexInfo, OLiA, SKOS-XL, and how those can be combined in the LOD cloud. Inspecting the BabelNet service that is using the *lemon* components. Introduction to LLOD formats for corpus annotation.

***Day 4**:* Working with concrete examples of (multilingual) dictionary, terminology and corpus-based lexical resources that can be mapped onto *lemon.*

***Day 5**:* Wrap-Up: finalizing the implementation and presentation by the participants of their work.

Expected level and prerequisites

This course is introductory in the sense that it does not presuppose experience in Semantic Web or Linked Data, but familiarity with linguistics and computation will be helpful.

Selected references

1. Hitzler, P., Krötzsch, M., & Rudolph, S. (2009). Foundations of Semantic Web Technologies. Chapman & Hall/CRC Textbooks in Computing.
2. Lexicon Model for Ontologies: Community Report, 10 May 2016, resulting from discussions and exercises conducted in the W3C Ontology-Lexica Community group. See https://www.w3.org/2016/05/ontolex/
3. Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.
4. McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. In Pustejovsky, J., & Wiebe, J. (eds). Language Resources and Evaluation, 46(4), 701-719, Berlin Heidelberg: Springer.
5. Bosque-Gil, J., Gracia, J., Cimiano, P., Stolk, S., Khan, F., Depuydt, K., de Does, J., Frontini, F., Kernerman, I. (2019). W3C OntoLex Lexicography Module Specification (lexicog). See https://www.w3.org/2019/09/lexicog/

Biography

Thierry Declerck is a Senior Consultant at the German Research Center for Artificial Intelligence (DFKI GmbH) since 1996, working in the field of multilingual language technologies. Thierry has acquired a larger number of projects in various fields, like multimedia semantics, information extraction, sentiment analysis (being the overall coordinator of the past European project "TrendMiner"), multilingual knowledge systems, etc. He is currently responsible for the DFKI contribution to the "Prêt-à-LLOD" project on "Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors" (https://pret-a-llod.github.io/). Thierry also contributes for the Austrian Center for Digital Humanities to the ELEXIS project (https://elex.is/), especially investigating the role of Linked Data for the representation and generation of lexical data. Additionally, Thierry is the Scientific Communication Manager of the COST Action CA18209 - European network for Web-centred linguistic data science ("NexusLinguarum", https://nexuslinguarum.eu/)

Julia Bosque-Gil is a postdoctoral researcher at the Distributed Information Systems Group at University of Zaragoza and a member of the Aragon Institute of Engineering Research (I3A). She has investigated the use of linguistic linked data for lexicography, and currently works on the representation and linking of multilingual resources as linguistic linked data as part of the Prêt-à-LLOD project. She co-leads NexusLinguarum COST Action Working Group 1 on Linked-data based Language Resources and co-chairs the Ontology-Lexica Community Group since October 2018.

SCHEDULE

Schedule 5– 9 July	Formal and Experimental Linguistics	Terminology and Lexicography	Grammar & Text
9:00am – 12:00pm	*Course 1*	D.J. Prinsloo	*Course 1*
2:00pm – 5:00pm	*Course 2*	Raquel Amaro & Rute Costa	*Course 2*
5:30pm – 8:30pm	*Course 3*	Thierry Declerck & Julia Bosque Gil	*Course 3*