STSM Topics

Active topics

Topic 1: Attitudes Detection and Representation in Survey Data

Outline & expected outcomes

Survey data provide a valuable source of information and research for different scientific disciplines. To study the information provided in survey data effectively, we need language tools allowing us to process them. The social sciences use case aims to build a library of tools that will enable the usage of survey data archives, organized according to linked data principles and generalizations about social attitudes clusters based on social media analysis and linking. First, we have to identify language phenomena such as discourse markers, attitude expressions, opinion detection, sentiment analysis, and then explore machine learning methods that detect and interpret them.

The goal of the proposed STSM is to exchange experience and know-how on using machine learning methods, such as transformer models and convolutional neural network for detecting and interpreting these language phenomena in texts from survey data, TED talks, and social media in English, Lithuanian, Hebrew, German, Macedonian. The outcome of the STSM will be an analysis of the results of a series of experiments over these three categories of datasets, and the extraction of semantic information, related to discourse markers, attitude expressions, opinion detection, and sentiment from them. Finally, the extracted semantic information will be analyzed and modeled according the LLOD principles.

The contribution of the proposed STSM to the scientific objectives of NexusLinguarum Action is that we will move forward and analyze one relevant use case for linguistic data science that will allow us to address some prototypical cases of interpretation of language phenomena with social significance. Further, we will exchange and derive best practices for semantic representation of cross-lingual linking and linguistic data analysis at a large scale.

Relevant WG(s)

WG4, WG1, WG2

Workplan
A series of STSMs are envisaged that will cover the following activities that are not interdependent:
1) constitution of corpora – collection selection, analysis, annotations, annotation schema
2) processing - familiarization with the software environment to run the models, constitution of experiment settings, running experiments, analyzing the results
3) defining a strategy for semantic representation of the language phenomena as linked data, deriving exemplary representations

Suggested duration
2 – 6 weeks

Contact person(s) and email
Dr Mariana Damova (mariana.damova@mozajka.co)

Host institution
Mozaika Ltd.
Solunska 52,
Sofia 1000,
Bulgaria
http://www.mozajka.co
Short Profile: Mozaika, The Humanizing Technologies Lab, provides research and development in the field of data science, natural interfaces (human-computer interaction), knowledge management and human insight. At Mozaika we are trying to leverage data science with natural interfaces to provide solutions tailored to human behavior, attitudes and comprehension. The company specializes in building information infrastructures that serve a variety of applications in data as a service or intelligence as a service modes. Our solutions are either human user facing or modules of larger systems. We cover verticals like business information delivery, human resources management, cultural heritage, water management, earth observation for AI, historical archives, content mapping, research infrastructures for the humanities, and intent to expand into Industry 4.0 and Smart Cities. The European Space Agency, the French Embassy in Bulgaria, Rittal GmbH, Ogylvy, Infeurope S.A., Sofia Municipality count amongst our clients. We partner with organizations like ESRI Bulgaria, SISTEMA GmbH, Gothenburg University, the Italian-Bulgarian Chamber of Commerce, Sofia University and the Bulgarian Academy of Sciences. All projects of Mozaika have technologically an innovative edge and abide to principles of human-centric design.

Topic 2: Modelling interlinear glosses for the web

Outline & expected outcomes
Interlinear glossed text represents a type of annotation frequently used in language documentation and in the humanities (esp., the philologies), as illustrated for an example from Armenian below:

Kat´ołikos-ə	*ut-um*	*ēr*
Catholicos-DEF	eat-IPFV	AUX.PST.3SG

'Catholicos was eating.'

The topic is to publish an IGT corpus (either provided by the guest or the host) in accordance with Linked Data principles and vocabularies. The host institution is involved in both the development of representation formalisms for interlinear glossed text (with the problem-specific vocabulary Ligt, based on the standard tools for the purpose, esp., Toolbox, FLEx and Xigt), but so far, this has not been connected with other standards for representing linguistic annotations on the web (esp., NIF, POWLA and Web Annotation). The topic and expected outcome of the research stay is to convert an existing IGT resource (preferably provided by the guest) to Ligt, and then to explore to what extent a mapping to more general RDF vocabularies such NIF, Web Annotation or POWLA requires extensions of the latter model(s). The expected outcome is a short whitepaper that will serve as direct input into on-going discussions in the W3C Community Group LD4LT.

Relevant WG(s)
Primarily WG1, T1.1 (data modelling) and T1.5 (IGT data represents a major part of linguistic data for low-resource languages), but also WG4, T4.2 (IGT data as used in the philologies) and T4.3 (IGT data as used in linguistics).

Workplan
(1) identification of candidate dataset(s) and target vocabularies (POWLA, NIF, Web Annotation) by guest and host
(2) introduction into Ligt by the host, optionally: introduction into POWLA, NIF, Web Annotation
(3) mapping the resource to Ligt and discussion of the results
(4) exploring possible mappings to selected target vocabulary in close collaboration with host
(5) short whitepaper summarizing challenges and achievements

Suggested duration and time period
8-12 weeks, start date flexible (to be coordinated with applicants)

Contact person(s) and email
Christian Chiarcos (chiarcos@informatik.uni-frankfurt.de), Max Ionov (ionov@cs.uni-frankfurt.de). *Please contact both of us*.

Host institution
Applied Computational Linguistics Lab
Goethe Universität Frankfurt
Robert-Mayer-Str. 11-15
60325 Frankfurt am Main, Germany
http://www.acoli.informatik.uni-frankfurt.de/

Topic 3: Concept alignment in linked dictionaries

Outline & expected outcomes
Translation inference across dictionaries is an established challenge in the LLOD context (see, e.g., https://tiad2019.unizar.es/). Within this STSM, we would like to explore an

extension of this problem from the bilingual to the multilingual use case: Instead of inferring translations between two languages, we aim to infer language-independent concepts with lexicalizations into multiple languages. The baseline for such an approach builds on the detection of translation cycles that allow to deduce multilingual concepts by collapsing bilingual translations („bank“@en – „riba“@ca – „orilla“@es – „bank“@en => *concept:bank1*; „bank“ @en – „banc“@ca – „banco“@es – „bank“@en => *concept:bank2*), resp. to provide alternative lexicalizations for these concepts („shore“@en – „riba“@ca – „orilla“@es – „shore“@en => „shore“@en – *concept:bank1*). The goal of the STSM is to implement, to evaluate, and, possibly, to improve this baseline method against different parts of the ACoLi Dictionary Graph (https://github.com/acoli-repo/acoli-dicts), the largest collection of machine-readable bilingual dictionaries currently available under an open license, currently covering more than 430 languages. The expected outcome is a baseline implementation and a whitepaper summarizing the implementation and possible improvements. Depending on the success of the baseline method, it can be used as a starting point to develop a future shared task on multilingual concept induction in the context of WG1.

Relevant WG(s)
Primarily WG1, T1.3 (data linking) and T1.5 (providing novel translation pairs for low-resource languages provided in the ACoLi Dictionary Graph).

Workplan
(1) identification of training and test data from the ACoLi Dictionary Graph or other datasets provided by the guest (1 week)
(2) implementation and evaluation of the baseline (3 weeks)
(3) iterative cycles of evaluation, refinement and discussion with the host (3 to 7 weeks)
(4) documentation (1 week)

Suggested duration and time period
8-12 weeks, start date flexible (to be coordinated with applicants)

Contact person(s) and email
Christian Chiarcos (chiarcos@informatik.uni-frankfurt.de), Jutta Nadland (j.nadland@em.uni-frankfurt.de). *Please contact both of us*.

Host institution:
Applied Computational Linguistics Lab
Goethe Universität Frankfurt
Robert-Mayer-Str. 11-15
60325 Frankfurt am Main, Germany
http://www.acoli.informatik.uni-frankfurt.de/

Topic 4: Aligning data categories for terminology management and language technology

Outline & expected outcomes
In the context of (Linguistic) Linked Open Data, the Ontology of Linguistic Annotations (OLiA, http://purl.org/olia/) represents the major terminology hub for linguistic categories (e.g.,

CommonNoun, etc.), similar to the role that DatCatInfo (http://datcatinfo.net/) serves for ISO standards such as TBX, LMF, etc. To facilitate synergies between both communities, their respective tools and resources, the goal is to explore the linking of DatCatInfo concepts with OLiA. Topic of the STSM is to familiarize the visiting researcher with OLiA specifics, introduce him/her into existing linkings with related resources (most notably ISOcat and the CLARIN Concept Registry), to work towards a conversion of DatCatInfo information to RDF and to provide a linking between DatCatInfo and OLiA by means of skos:broader, resp., rdf:type/rdfs:subClassOf relationships. The outcome will be a joint report about the linking effort and (to the extent feasible) a first version of the linking. Overall, this will contribute to establishing interoperability for language resoures in the terminology and language technology communities.

Relevant WG(s)
Primarily WG1 (esp. T1.1 data modelling), as a resource relevant to all WGs to the extent they address terminology and Linked Data (esp. WG2/T2.5, WG3/T3.3).

Workplan
(1) introduction to OLiA, OLiA linkings and applications by host
(2) selection and retrieval of DatCatInfo data with the help of its maintainers (contact person: Sue Ellen Wright, Kent State University) and host
(3) if necessary: introduction to ontology mapping by host
(4) development of a first draft mapping
(5) iterative refinement in close coordination with host and DCR maintainers
(6) short whitepaper summarizing challenges and achievements

Suggested duration and time period
8-12 weeks, start date flexible (to be coordinated with applicants)

Contact person(s) and email
Christian Chiarcos (chiarcos@informatik.uni-frankfurt.de), Jutta Nadland (j.nadland@em.uni-frankfurt.de). *Please contact both of us*.

Host institution
Applied Computational Linguistics Lab
Goethe Universität Frankfurt
Robert-Mayer-Str. 11-15
60325 Frankfurt am Main, Germany
http://www.acoli.informatik.uni-frankfurt.de/

Topic 5: Conversion of minority languages' resources into linguistic linked data

Outline and expected outcomes
The idea is that NexusLinguarum members interested in converting linguistic data of minority languages to LLOD visit a group with technical experience in such a process so the STSM applicant and the host team can analyse the data and plan and start its conversion

into RDF together. This is a valuable means to make such data interoperable and eventually to expand it by linking it to another linked data on the Web.

Relevant WG(s)
WG1 primarily

Relevance to NexusLinguarum objectives and workplan
Support of under-resourced languages is core for T1.5 in particular and a major goal for NexusLinguarum as a whole.

Suggested duration
minimum 2 weeks, ideally one to three months

Potential host institutions and contact details
STSMs on this topic can be carried out **on one of the following host institutions** depending on the type of resource(s) involved.

Host 1
Host: Distributed Informations Systems group (SID), University of Zaragoza, Spain
Type of resources: dictionaries, terminologies, glossaries, lexica; in any language but with preference to those that are present in Apertium (https://wiki.apertium.org/wiki/List_of_language_pairs)
Contact: Julia Bosque Gil (jbosque@unizar.es)
Short profile: The Distributed Information Systems research group (SID) belongs to the Aragon Institute for Engineering Research (I3A) - University of Zaragoza. The University of Zaragoza belongs to the list of Spanish universities included in the top 500 universities worldwide according to the Shanghai Ranking, and occupies the 406th position on the CWUR ranking. With the final goal of facilitating the interoperability between distributed, open, and dynamic information systems, our research group focuses on the development and use of semantic techniques for information retrieval, with emphasis on knowledge representation, ontologies, and Semantic Web services, among other areas. In linguistic data science research, we are particularly active in the transformation and linking of multilingual language resources to Linguistic Linked Data.
Location: C/ María de Luna nº 1, 50018 - Zaragoza, Spain.

Host 2
Host: CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy
Type of resources: dictionaries, lexica, glossaries, textual corpora
Contact: Marco C. Passarotti (marco.passarotti@unicatt.it)
Host: Laboratoire LIG, Université Grenoble Alpes, France
Type of resources: dictionaries ; with special interest in African Languages (tamajaq, haoussa, kanouri, zarma, fulfuldé/pulaar, wolof (Sénéga)l and nouchi (Côte d'Ivoire) ; also Berber, breton (Brittany) and na (Tibeto-Birman)
Contact: Gilles Sérasset (Gilles.Serasset@imag.fr)

Host 3
Host: Applied Computational Linguistics Lab, Goethe University Frankfurt, Germany
Type of resources: dictionaries, lexica, corpora
Contact: Christian Chiarcos (chiarcos@informatik.uni-frankfurt.de)

Host: Faculty of Information Technologies at the Czech Technical University in Prague, Czech Republic
Type of resources: lexicons, Wikipedia, DBpedia ; special interest in extraction of linguistic information for minority languages based on Wikipedia content
Contact: Milan Dojchinovski (milan.dojchinovski@fit.cvut.cz)

Host 4
Host: University of the Basque Country (UPV/EHU), Faculty of Arts, Spain
Type of resources: (historical) lexical data as LD; Basque
Contact: David Lindemann (david.lindemann@ehu.eus)
Short profile: At UPV/EHU Faculty of Arts, we are trying to set up a competence group for the digitization of Basque (historical) lexical resources. Basque is a minority language, which in the digital sphere is quite close to the state of the art, regarding NLP, but, at the moment, far behind with regard to the availability of digitized lexical resources. By our participation in actions of COST e-Lexicography and DARIAH WG "lexical resources" we have gained some skills, and we are now testing a digitization workflow that is scanned dictionary images - OCR - structure annotation as TEI XML - representation as linked data, looking also at compatibility with how lexical data is modeled in Wikidata. One reason for working with historical dictionaries is that due to their age they are public domain. Obviously, historical data is also interesting because of its extra challenges. We think that a consolidated competence group that could apply for larger research projects would be able to ask for the release of non-free lexical datasets, including up-to-date Basque dictionaries. We would definitely be interested in hosting research visits on the topic of (historical) lexical data as LD, and we could provide all necessary infrastructure for the meetings in our centre at Vitoria-Gasteiz.
Location: Micaela Portilla Research Center, Vitoria-Gasteiz, Spain