D2.1
Intermediate Activity Report
Working Group 2
“Linked data-aware NLP services”

Main authors:

John McCrae and Patricia Martín-Chozas

Project Acronym	NexusLinguarum
Project Title	European network for Web-centred linguistic data science
COST Action	18209
Starting Date	26 October 2019
Duration	48 months
Project Website	https://nexuslinguarum.eu/
Chair	Jorge Gracia
Main authors	John McCrae and Patricia Martín-Chozas
Contributors	Minna Tamper, Hugo Gonçalo Oliveira, Marco Maru, Mohammad Fazleh Elahi, Rute Costa, Elena Montiel-Ponsoda, Ciprian-Octavian Truică, Barbara McGillivray, Giedre Valunaite Oleskeviciene, Katerina Zdravkova, Sara Carvalho, Liudmila Mockienė, Jorge Gracia, Mariana Damova, Bharathi Raja Chakravarthi, Mike Rosner
Reviewer	NexusLinguarum core group team
Version | Status	Final
Date	29/11/2021

Acronyms List

CA	COST Action
EL	Entity Linking
ISO	International Organization for Standardization
KE	Knowledge Extraction
KM	Knowledge Management
LMF	Lexical Markup Framework
LD	Linked Data
LD4LT	Linked Data for Language Technology
LLD	Linguistic Linked Data
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LR	Language Resource
MT	Machine Translation
NLP	Natural Language Processing
QA	Question Answering
RDF	Resource Description Framework
SOTA	State Of The Art
STSM	Short Term Scientific Mission
SW	Semantic Web
TEI	Text Encoding Initiative
UC	Use Case
WG	Working Group
WSD	Word Sense Disambiguation

Table of Contents

Executive Summary

This report summarises the progress and future actions of Working Group 2 (WG2), “Linked data-aware NLP services”, as part of the NexusLinguarum COST Action (CA) CA18209. During the last months, we have successfully re-organised the leading structure of the group, which translates into fluent collaborations amongst the different group tasks and with other working groups. This document describes the work done as of November 2021, including task activities, interaction with other other working groups, STSMs and VMGs, and other events. The result of WG2 progress is materialised in the form of scientific publications, applications, and organization of datathons, workshops and conferences with the objective of finding the perfect symbiosis between LLOD and NLP.

1. Introduction

While Working Group 1 focuses on the generation, modeling and publication of language resources following the paradigm of Linked Open Data, Working Group 2 is in charge of using them to improve different Natural Language Processing tasks, such as Knowledge Extraction, Machine Translation, Question Answering, amongst others.

Therefore, the objectives of WG2 include the application of distributional models and neural networks in knowledge extraction; the exploration of relations amongst word embeddings and its conversion into Linked Data; the application of MT to generate language resources; the automatic translation of natural language into SPARQL queries and the application of LLOD resources to extract disambiguated knowledge, to mention but a few.

To fulfill those objectives, Working Group 2 is organized into five tasks:

1. **LLOD in Knowledge Extraction**, focused on the extraction of knowledge (structured information) from documents, with the help of neural networks and LD.
2. **LLOD in Machine Translation**, focused on the creation and application of LD for machine translation in under-resourced scenarios.
3. **LLOD in Multilingual Question Answering**, focused on the development of ontology-based multilingual QA.
4. **LLOD in Word Sense Disambiguation and Entity Linking,** focused on the creation of approaches and systems exploiting LD to automatically determine the meaning of an ambiguous word in context.
5. **LLOD in Terminology and Knowledge Management**, focused on the generation, modeling and application of terminologies in SW formats.

During the first months of the project, there was no fixed leadership structure for these tasks, but at this moment, each of the tasks is headed by one leader and, in some cases, by one co-leader, who are expert researchers in these areas (see Table 1). The WG2 mailing lists counts with 88 participants, from which around 25 regularly attend telcos.

Role	Name	Affiliation	Country
WG2 leader	Patricia Martín-Chozas	Universidad Politécnica de Madrid	Spain
WG2 co-leader	John P. McCrae	NUI Galway	Ireland
Task 2.1 leader	Hugo Gonçalo Oliveira	University of Coimbra	Portugal
Task 2.1 co-leader	Minna Tamper	Aalto University	Finland
Task 2.2 leader	Bharathi Raja Chakravarthi	NUI Galway	Ireland
Task 2.2 co-leader	John McCrae	NUI Galway	Ireland
Task 2.3 leader	Mohammad Fazleh Elahi	Uni Bielefeld	Germany
Task 2.3 co-leader	Philipp Cimiano	Uni Bielefeld	Germany
Task 2.4 leader	Marco Maru	Sapienza University of Rome	Italy
Task 2.5 leader	Elena Montiel-Ponsoda	Universidad Politécnica de Madrid	Spain
Task 2.5 co-leader	Rute Costa	Universidade Nova de Lisboa	Portugal

Table 1. WG2 Structure (as of November 1st, 2021).

2. Tasks Reports

2.1. Task 2.1 LLOD in Knowledge Extraction

Task Leaders:

- Leader: Hugo Gonçalo Oliveira
- Co-leader: Minna Tamper

General Overview

The task centers around themes of knowledge extraction from textual documents. Currently, its main focus is on exploring distributional models (e.g., word2vec, GloVe) and large neural language models (e.g., BERT, GPT) for the extraction of knowledge. This knowledge (i.e., named entities and / or relations) can be easily converted to LD and used in the creation or enrichment of new or existing knowledge bases.

Inspiration follows the acquisition of relations of different types from word embeddings (Drozd et al., 2016) and the utilization of neural language models as knowledge bases (Petroni et al., 2019). Also, by identifying relations and linguistic properties of words, the data can be used to create more detailed materials for training linguistic parsers and named entity extraction tools for highly inflected languages, such as Finnish.

Progress

- The HENKO ontology has been created for person names and one of its goals is to provide data for various knowledge extraction tools that can be used for different texts (Tamper et al., 2020). Also, initial systems have been created for knowledge extraction from Finnish language texts (Tamper et al., 2020) and their development continues. Here, the goal has been to create an ontology of information to enable identification of novel named entities from texts that can be otherwise disregarded by the machine learning based NER tools.
- GloVe embeddings were explored for the suggestion of lexico-semantic relations for enriching a Portuguese wordnet (de Paiva et al., 2012), taking advantage of analogy-solving methods (Gonçalo Oliveira et al., 2021).
- A BERT model pre-trained for Portuguese (Souza et al., 2020) was used as a masked language model and assessed in the automatic acquisition of lexico-semantic relations (e.g., hypernymy, part-of), by taking advantage of a set of handcrafted lexical patterns (Gonçalo Oliveira, 2021).
- Macedonian adjectives have a rich inflectional paradigm (Zdravkova and Petrovski, 2007). Language specific are the two different inflected forms for the adjectives with an identical lemma, root and morphosyntactic description. They depend on the etymology of the adjective and determine the prospective collocations and translation equivalents. Automatic extraction of the adjectives and collocation extraction has been finished, resulting in more than 300 adjectives and more than 2000 collocations. Further steps include the creation of the semantic components and the hierarchical network of the nouns that will unambiguously determine the exact adjective inflections.

Future activities

In the future we plan to organize or join online lunch seminars where we can share our work related to, for example, different tasks related to WG2 to share our experiences with other scholars. Currently, we are discussing with organizers of the DH Pizza Lunch Seminar to present work there. The DH pizza seminar has been organized online by Aalto University and University of Helsinki in Finland. We are also planning to collaborate with other working groups to write joint papers. In this scope, we are discussing the preparation of a survey on applications of Deep Learning and LD (T3.2), which should cover Knowledge Extraction.

2.2. Task 2.2 LLOD in Machine Translation

Task Leaders:

- Leader: Bharathi Raja Chakravarthi
- Co-leader: John McCrae

General Overview

LLOD in machine translation can have several applications in particular with respect to the collection of more data for under-resourced scenarios. As such, a particular focus of the work in this task is to do with the collection of novel resources for under-resourced languages. We hope that through the use of LLOD techniques we can build better resources that allow machine translation techniques to be applied more effectively and efficiently to new languages. A survey that charts the use of semantic web technologies in machine translation appears in D. Moussallem et al (2018).

Progress

- **TWB-Adapt project** - This project concerned the development of language resources for use by aid workers in the Rohingya refugee crisis. NUIG collaborated with Translators without Borders to develop novel translation resources, including some of the first digital resources for the Chittagonian and Rohingya languages.
- **DravidianLangTech Workshop** - Dravidian languages are primarily spoken in south India and Sri Lanka. Pockets of speakers are found in Nepal, Pakistan, Malaysia, Singapore, other parts of India and elsewhere in the world. We conducted a DravidianLangTech workshop at EACL 2021 to investigate challenges related to speech and language resource creation for Dravidian languages. We also conducted a shared task on machine translation in Dravidian languages. https://dravidianlangtech.github.io/2021/

Future activities

We plan to conduct another shared task on machine translation between Dravidian languages at DravidianLangTech-2022 at ACL 2022.

2.3. Task 2.3 LLOD in Multilingual Question Answering

Task Leaders:

- Leader: Mohammad Fazleh Elahi
- Co-leader: Philipp Cimiano

General Overview

The goal is to develop a model-based multilingual QA that uses an ontology lexicon in lemon format and automatically generates a lexicalized grammar used to interpret and parse questions into SPARQL queries. It is an alternative approach to machine learning technique that suffers from a lack of controllability, making the governance and incremental improvement of the system challenging, not to mention the initial effort of collecting and providing training data. The architecture consists of two components: (i) the grammar generator and (ii) the QA component. The grammar generator (Benz et al., 2020) takes a *lemon* lexicon as input and automatically creates lexicalized grammar rules. The QA component (Fazleh et al., 2021) is a web application that builds an efficient data structure to index the question data for later retrieval.

Progress

- **Multilingual question answering** - The grammar generation project generates about 1.8 million questions for English (Fazleh et al., 2021) from 336 lexical entries. It is further extended for German and Italian. The current version is capable of generating more than 1.6 million Italian questions (Nolano et al., 2021) over DBpedia.
- **Multilingual User interface** - The QA web application is extended for english, german, and italian. We conducted a usability test of the QA system prompting participants to answer specific questions and evaluate the performance. The evaluation was carried out with 161 students of the database course at Bielefeld University.

Future activities

Bangla is mainly spoken in Bangladesh and West Bengal (a state of India). The language is morphologically rich and contains complex syntactic constructions. We are developing a Bangla question answering system over WikiData.

We are planning to develop a QA system for Dravidian languages over WikiData. Dravidian languages are spoken in Sri Lanka, Pakistan, South India, Malaysia, and Singapore. It has official status in Sri Lanka, Tamil Nadu, Indian Union, Singapore. The focus is to extend the grammar generation project for Dravidian languages. The task is joint work with cooperation with NUIG, Indian, and Sri Lankan universities.

We are also extending the QA system for Spanish.

2.4. Task 2.4 LLOD in Word Sense Disambiguation and Entity Linking

Task Leaders:

- Leader: Marco Maru

General Overview

The application of LLOD is primary in the context of WSD and EL, both in supervised and unsupervised approaches. In fact a LLOD such as BabelNet (Navigli and Ponzetto, 2012) concurrently provide (i) a sound infrastructure to connect word senses across several languages by means of semantic relation edges, and (ii) a repository of knowledge that can be exploited to monitor system performance on traditional evaluation benchmarks.

Several works testify to the benefits of pivoting WSD methodologies on LLOD. Among those, authors have made use of contextualized sense embeddings exploiting LLOD to scale multilingually (Scarlini et al., 2020), whereas others have leveraged their structure to build test beds in low-resource environments (Pasini et al., 2021). More recently, the LLOD of BabelNet has also been successfully employed to aid a sense projection approach in creating high-quality training sets in multiple languages (Procopio et al., 2021).

In the context of Task 2.4, we aim to keep harnessing the wealth of knowledge encoded in sources of LLOD in order to further reduce the gap between English and other languages, as well as devising strategies to enhance the quality of the data they inherently feature.

Progress

So far, we started working on top of the efforts conducted by members of the Sapienza NLP group during the last two years, which already set a new state of the art for WSD applications, with results matching and surpassing the estimated human performance (Barba et al., 2021). The progress to report for this task can be summarized as follows:

- We submitted a research paper to a top-tier venue (under review) demonstrating that WSD is a task far from being considered solved. Particularly, we performed an extensive error analysis conducted on traditional evaluation benchmarks that use WordNet and proposed fresh test beds and challenge sets to better assess actual system capabilities;
- A sound experimental setup is being devised to (i) further delve into the use of Continuous Sense Comprehension (Barba et al., 2021) with a focus on cross-lingual settings, and (ii) to explore the synergies in multi-task approaches concurrently dealing with WSD and Semantic Role Labeling. In light of this, we aim to make further contributions available to the research community by submitting research papers during Q1 2022.

Future activities

Plans for Task 2.4 are mainly twofold: on the one hand, we aim to keep exploring unprecedented techniques to tackle WSD for instance the proper disambiguation and representation of discourse markers expressions in text and consequently have publications in top-tier venues featuring acknowledgements to the NexusLinguarum consortium. On the other hand, we intend to strengthen the dissemination of past and current works in the context of WSD and EL by re-activating the Sapienza NLP research team weekly reading group so as to host partners from all WGs and foster collaboration. At the same time, we look to involve authors of relevant WSD publications to take part in meetings and brainstorming sessions hosted by other research groups within interested WPs.

2.5. Task 2.5 LLOD in Terminology and Knowledge Management

Task Leaders:

- Leader: Elena Montiel-Ponsoda

- Co-leader: Rute Costa

General Overview

LLOD in terminology and knowledge management can serve several purposes. The representation formats provided by the LOD paradigm favor the integration of terminological resources previously isolated or difficult to discover and reuse. The fact that some linguistic and terminological resources are already provided in the LLOD cloud also contributes to the reuse of such resources in further NLP tasks. Terminological resources in LOD formats have proven their value and usefulness in knowledge management tasks, as models to structure and organise domain knowledge. In this regard, several technologies are emerging to cover gaps related to the creation and conversion of terminological resources into LOD formats. These technologies aim at speeding up the creation and exposition of terminological resources in such formats, and/or the conversion of traditional terminological resources into LOD resources. The final objective is to allow a more efficient use of terminological resources for its consumption by both humans and machines.

Progress

So far, several activities related to the main objectives of this task have been performed:

- **Terminology Summer School** organised by the NOVA CLUNL's board, with Rute Costa as one of the people responsible for the event, which took place in July 2021; Thierry Declerck, Julia Bosque Gil and Rute Costa participated as tutors in 2 different seminars)[1]
- **NexusLinguarum has sponsored the TOTH 2021 Conference** held in June 2021.[2]
- During the months of July to September, an **exchange visit** (see section 3 for STSMs) happened between researchers at UPM (Patricia Martín-Chozas) and DFKI (Thierry Declerck). Patricia Martín-Chozas visited DFKI in Saarbrücken to work on the TermLex model, an extension of the modelling mechanisms foreseen in the current version of the Ontolex model to deal with terminological data. The result of this effort has been proposed to the Ontology-Lexicon Community Group of the W3C for further discussion and approval by the community.
- Patricia Martín-Chozas presented **TermLex at the Ontolex workshop**[3], held on 4th September in Zaragoza, Spain, co-located with the LDK 2021 international conference.
- Workshop "**Terminology in the 21st century: many faces, many places (Term21)**" to be co-located with LREC 2022 [proposal submitted]

Future activities

In the framework of this task, we are planning the edition of a state of the art paper in "LOD in Terminology and Knowledge Management" as a joint effort, to provide a thorough survey of resources, technological solutions, and projects around this topic.

[1] https://clunl.fcsh.unl.pt/lisbon-summer-school-in-linguistics-2021/

[2] http://toth.condillac.org/conference

[3] http://2021.ldk-conf.org/post-conference-w3c-day/

In the specific issue of terminology modelling, we intend to propose the organisation of a tutorial that would be co-located with a relevant conference in the area (LREC, LDK, EKAW, Coling, FOIS, K-CAP, EADH, etc., still to be decided). We have identified a growing demand on modelling solutions for terminological resources, and count on experts in ISO standards and Ontolex among the participants in this task.

3. Short Time Scientific Missions and Virtual Mobility Grants

UPM-DFKI STSM: During this 3-month STSM, Patricia Martín-Chozas from Universidad Politécnica de Madrid (UPM) visited the German Research Center for Artificial Intelligence (*Deutsches Forschungszentrum für Künstliche Intelligenz*, DFKI) to work with Thierry Declerck. This STSM was motivated by the prior cooperation of the applicant group, the Ontology Engineering Group from UPM, with the Speech and Language Technology lab of DFKI GmbH (Berlin) in the context of the Lynx project and the current cooperation with the Multilinguality and Language Technology lab of DFKI GmbH (Saarbrücken, which is the hosting institution for the STSM) in the context of the Prêt-à-LLOD project. Such collaboration has demonstrated the affinity of the work performed in both institutions, which, in this case, lies in the representation of domain-specific language resources (terminologies), that puts together the work carried out in WG1 and WG2. The most important output of this STSM is a draft specification of the possible terminology extension module for OntoLex-Lemon to be considered by the W3C “Ontology Lexica” community group, which is available in the community group Wiki site.

4. Organised Events

4.1 Bielefeld's Hackathon

The multilingual question answering (Fazleh et al., 2021) uses an ontology lexicon in the *lemon* format and automatically generates a lexicalized grammar that can be used to interpret and parse questions into SPARQL queries. The approach gives maximum control over the QA system to the developer of the system as every lexicon entry added to the lexicon increases the coverage of the grammar, and thus of the QA system, in a predictable way.

The hackathon aims to improve the QA system on the combination of semantic web and language technologies. The event was online and more than 12 participants from different backgrounds (computer science, linguistics, computational linguistics, semantic web, etc.) actively participated in the hackathon. The participants include academia, industry, and freelancers. An easy-to-use recipe[4] was provided to install and run the grammar generation and user interface. The participants improve the QA system as follow:

- The coverage of the grammar is extended for multiple languages.
- The grammar generation task is extended to support Italian.
- the QA system is ported to other datasets such as Wikidata and ArCo (Italian culture heritage linked data)
- The user interface is extended to support multilingual questions and multiple datasets.

The activities and the outcome of the hackathon with code and results are available on the webpage[5]. The hackathon led to cooperation with NLP Research Group, University of Naples "L'Orientale", Italy. We developed the first and largest question answering system over linked data for Italian. The outcome titled “An Italian Question Answering System based on grammars automatically generated from ontology lexica” (Nolano et al., 2021) is accepted in the 8th Italian conference of computational linguistics.

4.2 Lisbon Training School

The Lisbon Summer School took place at the Universidade NOVA de Lisboa from 5-7 July 2021[6]. In the 2nd area of the Summer School (Terminology and Lexicography), and under the support of NexusLinguarum, one of the courses was entitled “Introduction to Linked Open Data in Linguistics”, with Thierry Declerck (NexusLinguarum's Science Communication Officer and WG3 co-leader) and Julia Bosque-Gil (WG1 co-leader until October 2021) as lecturers. A second seminar entitled “Linguistic Digital Data in Terminology and

[4] https://scdemo.techfak.uni-bielefeld.de/qahackathon/tutorial/coverage.php
[5] https://scdemo.techfak.uni-bielefeld.de/qahackathon/index.php
[6] https://clunl.fcsh.unl.pt/en/lisbon-summer-school-in-linguistics-2021/

Lexicography" was given by Raquel Amaro and Rute Costa (WG2 co-leader of the Task 2.5). With more than 45 participants for each seminar, both onsite and online, the course aimed to provide people in the fields of digital humanities and computational linguistics with the theoretical underpinnings, as well as practical skills, in the topics of linked data and semantic technologies as applied to linguistics and lexical data.

4.3 TOTh Conference

The TOTh Conference 2021[7] was held on June 3-4 2021, both onsite and online, with the purpose of bringing together researchers, professionals and, more generally, all those interested in issues related to language and knowledge engineering. With the support of NexusLinguarum, among other institutions, this conference had more than 80 participants, and included talks by Thierry Declerck (NexusLinguarum's Science Communication Officer and WG3 co-leader), as well as by Rute Costa (Task 2.5 co-leader) and Sara Carvalho (WG4 leader).

[7] http://toth.condillac.org/wp-content/uploads/2021/04/TOTh_2021_Final_Program_Online_En.pdf

5. Interaction with other WGs

WG2-WG1

The STSM reported in Section 3, between UPM and DFKI members, merges the objectives of Task 2.5 (LLOD in terminology and knowledge management) and Task 1.1 (LLOD modelling), since the work performed during this 3-month STSM was focused on the design of a new vocabulary extension to represent terminologies as Linked Data. This extension proposal is planned to be discussed within the Ontology Lexica Community Group during the following months in order to be officially approved by the group chairs.
Also in the intersection of WG1 and WG2, a work on cross-lingual model transfer in the Pharmaceutical domain was carried out by members of the Action (Gracia et al., 2020). This includes the transformation of a new version of the Apertium dictionary data in RDF (WG1) as well as its use to support LD-aware NLP services such as sentiment analysis (WG2).

WG2-WG4

The WG2-WG4 collaboration is twofold:

1. A humanities case study (Armaselu et al., 2021a) that proposes an interdisciplinary approach including methods from disciplines such as the history of concepts, linguistics, natural language processing (NLP), and Semantic Web was published at The 3rd Conference on Language, Data and Knowledge (LDK2021). The focus of the paper is to create a comparative framework for detecting semantic change in multilingual historical corpora and generating diachronic ontologies as linguistic linked open data (LLOD). Thus, the paper (1) explores emerging trends in knowledge extraction, analysis, and representation from linguistic data science and (2) describes the main elements of the methodological framework and preliminary planning of the intended workflow.
2. A survey (Armaselu et al., 2021b) that presents an overview of the LL(O)D and NLP methods and tools for detecting and representing semantic change was accepted for publication at the Semantic Web Journal. The focus of the paper is how to use such methods and tools in humanities research. The final aim is to provide the starting points for the construction of a workflow and set of multilingual diachronic ontologies within the humanities. Furthermore, the survey focuses on the essential aspects needed to understand the current trends and to build applications in this area of study.

Additionally, the first workshop on Sentiment Analysis & Linguistic Linked Data (SALLD-1, see https://www.salld.org/), took place on September 1, 2021, and was co-located with the

LDK 2021 – 3rd Conference on Language, Data and Knowledge, in Zaragoza, Spain. It was organised by members of WG4 and WG2.

WG2-WG3-WG4

The STSM “Researching discourse markers expressing opinion with machine learning techniques in a multilingual corpus” (by Giedre Valunaite Oleskeviciene), held during the period from 16/08/2020 to 30/08/2020, was a breakthrough point in the use case 4.2.2 (Social Sciences) in collaboration with WG3 and WG2.

The purpose of the STSM was providing linguistic processing for several languages analyzing multilingual corpus data in English, Bulgarian and Lithuanian, in preparation of the data for the research of automatic detection of discourse markers expressing opinion by using machine learning.

First, we enriched TED-EHL parallel corpus-based social media texts with 4 languages so that the multilingual corpus contains alignments of Lithuanian, Bulgarian, Hebrew, Portuguese, Macedonian, and German languages, with English as pivot language, and with a size of 1.3 million sentences. Then the part of the enriched multilingual corpus comprising 2428 English-Bulgarian-Lithuanian aligned sentences containing the multiword expressions (MWE) as discourse markers or content expressions was manually annotated (1 or 0) in preparation for the machine learning experiments. The manual annotation of the data was carried out in order to refine the data for the successive elaboration, aiming to reach automatic detection of discourse markers expressing opinion. Therefore, during the STSM, we produced the gold standard which was later applied for machine learning.

The STSM “Corpus Analysis of Covid and health-related metaphors” (by Liudmila Mockienė), held during the period from 16/08/2020 to 30/08/2020, was carried out within the research conducted by Kristina Štrkalj Despot, Ana Ostroški Anić and Petya Osenova, who work on a Use Case of Public Health (4.4.1) under Task 4.4 Life Sciences in WG4, also with connections to WG2 and WG3. The aim of the STSM was the extension of multilingual resources, i.e. carrying out research on COVID-19 and health-related metaphors in Lithuanian, as an under-resourced language, based on a multilingual corpus ParlaMint-LT 2.0 (Lithuanian parliamentary corpus). The greatest benefit of the STSM was the possibility to combine forces and not only analyse and process the data, but also represent it as an initial hierarchy/ontology of the frames obtained together with the related entries, which were converted into an interoperable format as LOD, ready to be incorporated into other resources. This goal was possible to achieve only due to close collaboration with the host institution. The STSM also targeted discussions on the possibilities of further joint analysis of tendencies of the use and spread of COVID-19 and health-related metaphors in other languages and domains (news media and social media in a cross-lingual setting), focusing on the Lithuanian data, which will enable both comparison of the research results with the data from ParlaMint-LT 2.0 and the data in other languages.

6. Future Directions

Next steps in WG2 include the organisation of events as a meeting point for researchers from different tasks and working groups, such as the DH Pizza Lunch Seminar of Aalto University; the shared task on machine translation at DravidianLangTech-2022; the weekly reading group of the Sapienza University to foster collaboration with other WGs; the organisation of a tutorial for terminology modeling that would be co-located to a relevant conference in the area, and the organisation of the 4th Summer Datathon on LLOD. Such a datathon is planned in Cercedilla (Spain) in June 2022 and will constitute the second training school of NexusLinguarum. This will be a continuation of the fist training school that took place online in February 2021 (EUROLAN'21) and will focused on WG1 and WG2 developments.

Furthermore, there are several envisioned collaborations with WG3, specifically T3.2 Deep learning and neural approaches for linguistic data, given its relevance for some of the WG2 tasks. This may include a joint survey-like paper or the joint organisation of a workshop (e.g., a second edition of the Workshop on Deep Learning and Neural Approaches for Linguistic Data); and a state of the art paper on “LOD in Terminology and Knowledge Management” and the intersection with the previously mentioned techniques.

7. WG2 Related Publications

Gonçalo Oliveira, Hugo, Aguiar, Fredson, and Rademaker, Alexandre (2021). On the Utility of Word Embeddings for Enriching OpenWordNet-PT. In Proceedings of 3rd Conference on Language, Data and Knowledge (LDK 2021), volume 93 of OASIcs, pages 21:1–21:13, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Gonçalo Oliveira, Hugo (2021). Acquiring Lexico-Semantic Knowledge from a Portuguese Masked Language Model (Extended Abstract). Workshop on Deep Learning and Neural Approaches for Linguistic Data. Skopje, North Macedonia.

Viktoria Benz, Philipp Cimiano, Mohammad Fazleh Elahi, Basil Ell (2020), Generating Grammars from lemon lexica for Questions Answering over Linked Data: a Preliminary Analysis. *In: NLIWOD workshop at ISWC. vol. 2722, pp. 40–55. CEUR-WS.org.*

Mohammad Fazleh Elahi, Basil Ell, FrankGrimm, and Philipp Cimiano (2021). QuestionAnswering on RDF Data based on GrammarsAutomatically Generated from Lemon Models. *In SEMANTiCS Conference, Posters and Demonstrations.*

Gennaro Nolano, Mohammad Fazleh Elahi, Maria Pia di Buono, Basil Ell, Philipp Cimiano. (2021). An Italian Question Answering System based on grammars automatically generated from ontology lexica. Eighth Italian Conference on Computational Linguistics.

Oleškevičienė, Giedrė Valūnaitė, Liebeskind, Chaya, Trajanov, Dimitar, Silvano, Purificação, Chiarcos, Christian & Damova, Mariana (2021). Speaker Attitudes Detection through Discourse Markers Analysis. In Garabík (ed.) Book of Abstract of the Workshop on Deep Learning and Neural Approaches for Linguistic Data, Skopje, North Macedonia, URL: https://www.juls.savba.sk/attachments/workshop_20210930_en/workshop_20210930_book_of_abstracts.pdf.

Oleškevičienė, Giedrė Valūnaitė & Liebeskind, Chaya (2021). Multiword expressions as discourse markers in Hebrew and Lithuanian. In Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age (pp. 46-56). https://aclanthology.org/2021.motra-1.5/

Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Andrius Utka, Giedrė Valūnaitė Oleškevičienė, Marieke van Erp. *LL(O)D and NLP Perspectives on Semantic Change for Humanities Research*. Semantic Web Journal (accepted for publication 2021). URL: link

Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė (2021). *HISTORIAE, HIStory of culTural transfORmatIon as linguistIc dAta sciEnce. A Humanities Use Case.* Conference on Language, Data and Knowledge (LDK2021). DOI: 10.4230/OASIcs.LDK.2021.34

Jorge Gracia, Christian Fäth, Matthias Hartung, Max Ionov, Julia Bosque-Gil, Susana Veríssimo, Christian Chiarcos, Matthias Orlikowski (2020). *Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain*, in Proc. of 19th International Semantic Web Conference (ISWC 2020), pp. 499–514.

References

Barba, E., Procopio, L., & Navigli, R. (2021). ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1492-1503.

Drozd, A., Gladkova, A., and Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In Proceedings of 26th International Conference on Computational Linguistics: Technical Papers, pages 3519–3530, Osaka, Japan. COLING 2016 Organizing Committee.

Minna Tamper, Petri Leskinen, Jouni Tuominen and Eero Hyvönen (2020). Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology. *3rd Workshop on Humanities in the Semantic Web (WHiSe 2020)*, pp. 3-14, CEUR Workshop Proceedings, vol. 2695.

Minna Tamper, Arttu Oksanen, Jouni Tuominen, Aki Hietanen and Eero Hyvönen (2020). Automatic Annotation Service APPI: Named Entity Linking in Legal Domain. *The Semantic Web: ESWC 2020 Satellite Events* (Harth, Andreas, Presutti, Valentina, Troncy, Raphaël, Acosta, Maribel, Polleres, Axel, Fernández, Javier D., Xavier Parreira, Josiane, Hartig, Olaf, Hose, Katja and Cochez, Michael (eds.)), Lecture Notes in Computer Science, vol. 12124, pp. 208-213, Springer-Verlag.

D. Moussallem, M. Wauer, A.N. Ngomo, Machine translation using semantic web technologies: A Survey, Web Semantics: Science, Services and Agents on the World Wide Web (2018), https://doi.org/10.1016/j.websem.2018.07.001

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, 193, 217-250.

Pasini, T., Raganato, A., & Navigli, R. (2021). XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. In Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press.

Petroni, F., Rocktaschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473. ACL.

Procopio, L., Barba, E., Martelli, F., & Navigli, R. (2021). MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), pages 3915-3921.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In Proceedings of 24th International Conference on Computational Linguistics, COLING (Demo Paper).

Scarlini, B., Pasini, T., & Navigli, R. (2020). With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3528-3539.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In Proc of Brazilian Conf on Intelligent Systems (BRACIS 2020), volume 12319 of LNCS, pages 403–417. Springer.

Zdravkova, K., & Petrovski, A. (2007). Derivation of Macedonian verbal adjectives. In Proceedings of international conference Recent advances in natural language processing (RANLP), pages 661-665.