Nexus Linguarum

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

D2.3
Final Activity Report
Working Group 2
“Linked data-aware NLP services”

Main authors:

Patricia Martín Chozas and Andon Tchechmedjiev

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	Patricia Martín Chozas and Andon Tchechmedjiev
Contributors	Patricia Martín Chozas, Andon Tchechmedjiev, Barbara McGillivray, Simone Tedeschi, Lucía Pitarch, Hugo Gonçalo Oliveira, Mohammad Fazleh Elahi, Simone Tedeschi, Sara Carvalho, Rute Costa, Ricardo Rodrigues, Cremaschi Marco
Reviewer	NexusLinguarum core group team
Version | Status	Final
Date	22/4/24

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 of the second and final period of Working Group 2 (WG2), “Linked data-aware NLP services”, as part of the NexusLinguarum COST Action (CA)

CA18209. During the last months, the structure of the working group has been reorganised and the number of participants has grown, enabling fluent collaborations amongst the different group tasks and with other working groups. This document describes the work done as of April 2024, 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 organisation 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, modelling 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 fulfil those objectives, Working Group 2 is organised 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, modelling 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 127 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	Andon Tchechmedjiev	IMT Mines Alès	France
Task 2.1 leader	Hugo Gonçalo Oliveira	University of Coimbra	Portugal
Task 2.1 co-leader	Lucía Pitarch	Universidad de Zaragoza	Spain
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	Simone Tedeschi	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 April, 2024).

2. Tasks Reports

2.1. Task 2.1 LLOD in Knowledge Extraction

Task Leaders:

- Leader: Hugo Gonçalo Oliveira
- Co-leader: Lucía Pitarch

General Overview

The task centres 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 utilisation 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

- **Multilingual Lexical Relation Acquisition with BATS manual translation for cross-lingual transference:**
  - **Title**: MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations
  - **Authors:** Dagmar Gromann, Hugo Goncalo Oliveira, Lucia Pitarch, Elena-Simona Apostol, Jordi Bernad, Eliot Bytya, Chiara Cantone, Sara Carvalho, Francesca Frontini, Radovan Garabik, Jorge Gracia, Letizia Granata, Fahad Khan, Timotej Knez, Penny Labropoulou, Chaya Liebeskind, Maria Pia di Buono, Ana Ostrovski Ana, Sigita Rackeviena, Ricardo Rodrigues, Gilles Sérasset, Linas Selmistraitis, Mahammadou Sidibé, Purificaçao Silvano, Blerina Spahiu, Enriketa Sogutlu, Ranka Stankova, Ciprian-Octavian Truica, Giedre Valanaite, Slavko Zitnik and Katerina Zdravkova
  - **Abstract**: Understanding the relation between the meanings of words is an important part of comprehending natural language. Prior work has either focused on analysing lexical semantic relations in word embeddings or probing pretrained language models (PLMs), with some exceptions. Given the rarity of highly multilingual benchmarks, it is unclear to what extent PLMs capture relational knowledge and are able to transfer it across languages. To start addressing this question, we propose MultiLexBATS, a multilingual parallel dataset of lexical semantic relations adapted from BATS in 15 languages including low-resource languages, such as Bambara, Lithuanian, and Albanian. As experiment on cross-lingual transfer of relational knowledge, we test the PLMs' ability to (1) capture analogies across languages, and (2) predict translation targets. We find considerable differences across

relation types and languages with a clear preference for hypernymy and antonymy as well as romance languages.
    - **Venue:** LREC-COLING 2024
- **Multilingual Lexical Relation Acquisition with conceptual grounding using multilingual verbalizations in PTLMs**
    - **Title:** Building MUSCLE, a Dataset for MUltilingual Semantic Classification of Links between Entities
    - **Authors:** Lucia Pitarch, Carlos Bobed, David Avián, Jorge Gracia, Jorge Bernad
    - **Abstract:** In this paper we introduce MUSCLE, a dataset for MUltilingual lexico-Semantic Classification of Links between Entities. The MUSCLE dataset was designed to train and evaluate Lexical Relation Classification (LRC) systems with 27K pairs of universal concepts selected from Wikidata, a large and highly multilingual factual Knowledge Graph (KG). Each pair of concepts includes lexical forms in 25 languages and is labelled with up to five possible lexico-semantic relations between the concepts: hypernymy, hyponymy, meronymy, holonymy, and antonymy. Inspired by Semantic Map theory, the dataset bridges lexical and conceptual semantics, is more challenging and robust than previous datasets for LRC, avoids lexical memorization, is domain-balanced across entities, and enables enrichment and hierarchical information retrieval.
    - **Venue:** LREC-COLING 2024
- **Lexical Relation Extraction with PTLMs for Portuguese**
    - **Title:** BATS-PT: Assessing Portuguese Masked Language Models in Lexico-Semantic Analogy Solving and Relation Completion
    - **Authors:** Hugo Gonçalo Oliveira, Ricardo Rodrigues, Bruno Ferreira, Purificação Silvano and Sara Carvalho.
    - **Conference:** PROPOR 2024
    - **Abstract:** This paper presents BATS-PT, the manual translation of the lexicographic portion of the Bigger Analogy Test Set (BATS) to European Portuguese. BATS-PT covers ten types of lexicosemantic analogies and can be used for assessing word embeddings and language models. Following this, the dataset is showcased while assessing two pretrained language models for Portuguese, BERTimbau and Albertina, in two tasks: analogy solving and relation completion, both in zero- and few-shot mask-prediction approaches. Experiments reveal different performance across relations and, in both tasks, the best overall performance was achieved with BERTimbau, in a five-shot scenario. We further discuss the limitations of the reported experiments and directions towards future improvements in these tasks.
- **Lexical Relation Extraction with GPT3 for Portuguese**
    - **Title:** GPT3 as a Lexical Knowledge Base for Portuguese?
    - **Authors**: Hugo Gonçalo Oliveira and Ricardo Rodrigues
    - **Abstract:** We test the GPT3 language model in zero- and few-shot acquisition of lexico-semantic knowledge in Portuguese, with simple instruction prompts, and compare it with a BERT-based approach. Results are assessed in two test sets: TALES and the Portuguese translation of BATS. GPT3 outperforms BERT in all relations, with the few-shot approach being the best overall and for the majority of relations. Scores in both datasets further suggest that, despite their different creation approaches, they are equally suitable for this kind of evaluation.

- **Venue:** LDK 2023

- **Minimal prompting for Lexical Relation Acquisition**
  - **Title:** No clues, good clues: Out of context Lexical Relation Classification
  - **Authors:** Lucía Pitarch, Jorge Bernad, Licri Dranca, Carlos Bobed Lisbona, and Jorge Gracia
  - **Abstract:** The accurate prediction of lexical relations between words is a challenging task in Natural Language Processing (NLP). The most recent advances in this direction come with the use of pre-trained language models (PTLMs). A PTLM typically needs ``well-formed" verbalised text to interact with it, either to fine-tune it or to exploit it. However, there are indications that commonly used PTLMs already encode enough linguistic knowledge to allow the use of minimal (or none) textual context for some linguistically motivated tasks, thus notably reducing human effort, the need for data pre-processing, and favouring techniques that are language neutral since do not rely on syntactic structures. In this work, we explore this idea for the tasks of lexical relation classification (LRC) and graded Lexical Entailment (LE). After fine-tuning PTLMs for LRC with different verbalizations, our evaluation results show that very simple prompts are competitive for LRC and significantly outperform graded LE SoTA. In order to gain a better insight into this phenomenon, we perform a number of quantitative statistical analyses on the results, as well as a qualitative visual exploration based on embedding projections.
  - **Venue:** ACL 2023
- **Knowledge Extraction for Portuguese:**
  - **Title:** On the Acquisition of WordNet Relations in Portuguese from Pretrained Masked Language Models
  - **Authors:** Hugo Gonçalo Oliveira
  - **Abstract:** This paper studies the application of pretrained BERT in the acquisition of synonyms, antonyms, hypernyms and hyponyms in Portuguese. Masked patterns indicating those relations were compiled with the help of a service for validating semantic relations, and then used for prompting three pretrained BERT models, one multilingual and two for Portuguese (base and large). Predictions for the masks were evaluated in two different test sets. Results achieved by the monolingual models are interesting enough for considering these models as a source for enriching wordnets, especially when predicting hypernyms of nouns. Previously reported performances on prediction were improved with new patterns and with the large model. When it comes to selecting the related word from a set of four options, performance is even better, but not enough for outperforming the selection of the most similar word, as computed with static word embeddings.
  - **Venue:** Global WordNet Conference 2023
- **Comparative analysis of Serbian and Macedonian by Katerina and Ranka**
- **Inflectional Ambiguity of Macedonian Adjectives:**
  - **Title:** Resolving Inflectional Ambiguity of Macedonian Adjectives
  - **Author:** Katerina Zdravkova
  - **Abstract:** Macedonian adjectives are inflected for gender, number, definiteness and degree, with in average 47.98 inflections per headword. The inflection paradigm of qualificative adjectives is even richer, embracing 56.27 morphophonemic alterations. Depending on the word they were derived from, more than 600 Macedonian adjectives have an identical headword and two

different word forms for each grammatical category. While non-verbal adjectives alter the root before adding the inflectional suffixes, suffixes of verbal adjectives are added directly to the root. In parallel with the morphological differences, both types of adjectives have a different translation, depending on the category of the words they have been derived from. Nouns that collocate with these adjectives are mutually disjunctive, enabling the resolution of inflectional ambiguity. They are organised as a lexical taxonomy, created using hierarchical divisive clustering. If embedded in the future spell-checking applications, this taxonomy will significantly reduce the risk of forming incorrect inflections, which frequently occur in the daily news and more often in the advertisements and social media.
    - **Venue:** Globalex Workshop on Linked Lexicography at LREC 2022
- **Metaphor extraction:**
    - **Title:** MEAN: Metaphoric Erroneous ANalogies dataset for PTLMs metaphor knowledge probing
    - **Authors:** Lucía Pitarch, Jorge Bernad and Jorge Gracia
    - **Abstract:** Despite significant progress obtained in Natural Language Processing tasks thanks to Pre-Trained Language Models (PTLMs), figurative knowledge remains a challenging issue. This research sets a milestone towards understanding how PTLMs learn metaphoric knowledge by providing a novel hand-crafted dataset, with metaphorical analogy pairs where per correct analogy pair, other three erroneous ones are added controlling for the semantic domain and the semantic attribute. After using our dataset to fine-tune SoTa PTLMs for the multiclass classification task we saw that they were able to choose the correct term to fit the metaphor analogy around the 80\% of the times. Moreover, thanks to the added erroneous examples on the dataset we could study what kind of semantic mistakes the model was making.
    - **Venue:** LDK 2023
- **Analyses of Networks of Politicians Based on Linked Data**
    - **Title:** Analyses of Networks of Politicians Based on Linked Data: Case ParliamentSampo
    - **Authors**: Henna Poikkimäki, Petri Leskinen, Minna Tamper and Eero Hyvönen
    - **Abstract**: In parliamentary debates the speakers make reference to each other. By extracting and linking named entities from the speeches it is possible to construct reference networks and use them for analysing networks of politicians and parties and their debates. This paper presents how such a network can be constructed automatically, based on a speech corpus 2015–2022 of the Parliament of Finland, and be used as a basis for network analysis.
    - **Venue**: Semantic Web and Ontology Design for Cultural Heritage (SWODCH 2022)
- Research Stay at Vienna in October 2023 by Lucía Pitarch and Hugo Gonçalo Oliveira for Multilingual Lexical Relation Acquisition
- Research Visit at Zaragoza by Hugo Gonçalo Oliveira

Future activities

In the future we plan to continue exploring the reasoning abilities of Large Language Models towards extracting knowledge, especially semantic and lexical knowledge. To do so, we aim to structure the datasets provided in the deliverable as Linked Data so that they can be linked to other resources and further explore how Knowledge Graphs and the injection of relevant information can shed more light on the information encoded in Language Models and improve their robustness and performance.

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. The focus of this task was on the use of LLOD techniques to build better resources that allow machine translation techniques to be applied more effectively and efficiently to new languages. The progress reported was performed in the two first years of the action. Afterwards, the progress on this task stopped, since more efforts were made in other tasks.

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/

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 (Elahi et al., 2024) 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.

Based on preliminary research work (Benz et al., 2020), this thesis presents a multilingual question-answering (QA) approach (QueGG-Multi) that relies on a model of a lexicon-ontology interface in which the meaning of lexical entries is specified with respect to a given vocabulary or ontology. More specifically, given a lexicon in a particular language and for a particular vocabulary, the approach automatically generates a grammar (Elahi et al., 2021) that allows questions in the specific language to be parsed into SPARQL queries. The research shows that this approach outperforms current machine-learning-based QA approaches on QALD benchmarks. Furthermore, it demonstrates the extensibility of the approach to different languages by adapting it to German, Italian (Nolano et al., 2021), and Spanish.

While the approach provides state-of-the-art performance on benchmarks, a crucial prerequisite for the grammar generation approach is the availability of a Lemon lexicon that describes which lexical entries can be used to verbalise the elements of a particular dataset in a specific language. To this end, the work shows (Elahi et al., 2023) that such a lexicon can be induced automatically to some extent using the LexExMachina approach that builds on association rules to find correspondences between lexical elements and ontological vocabulary elements. The results show that the method could be used to bootstrap the semi-automatic creation of a lexicon, thus having the potential to significantly reduce the human effort involved in producing lexicons for languages beyond English.

Progress

- **Multilingual Question Answering over Linked Data**
  - **Title:** Multilingual Question Answering over Knowledge Graphs building on a model of the lexicon-ontology interface
  - **Authors:** Mohammad Fazleh Elahi, Basil Ell, Gennaro Nolano, Phillip Ciamiano
  - **Abstract:** Towards the development of QA systems that can be ported across languages in a principled manner without the need of training data and towards systems that can be incrementally adapted and improved after deployment, we follow a model-based approach to QA that supports the extension of the lexical and multilingual coverage of a system in a declarative manner. The approach builds on a declarative model of the lexicon ontology interface, OntoLex lemon, which enables to specify the meaning of lexical entries with respect to the vocabulary of a particular dataset. From a specific

lexicon describing the meaning of lexical entries with respect to a given vocabulary, in our approach a QA grammar can be automatically generated that can be used to parse questions into SPARQL queries. We show that this approach outperforms current QA approaches on the QALD benchmarks. Furthermore, we demonstrate the extensibility of the approach to different languages by adapting it to German, Italian, and Spanish. We evaluate the approach with respect to the QALD benchmarks on five editions (i.e., QALD-9, QALD-7, QALD-6, QALD-5, QALD-3) and show that state-of-the-art results can be achieved in both the training and test data. So far there is no system described in the literature that works for at least 4 languages while reaching state-of-the-art performance on all of them. Finally, we demonstrate the low efforts necessary to port the system to a new dataset and vocabulary.
    - **Venue:** Semantic Web Journal
- **Automatic Lexicon Induction for Multilingual Question Answering System**
    - **Title:** A framework for the automatic induction of ontology lexica for Question Answering over Linked Data
    - **Authors:** Mohammad Fazleh Elahi, Basil Ell, Phillip Ciamiano
    - **Abstract:** An open issue for Semantic Question Answering Systems is bridging the so called lexical gap, referring to the fact that the vocabulary used by users in framing the question needs to be interpreted with respect to the logical vocabulary used in the data model of a given knowledge base or knowledge graph. Building on previous work to automatically induce ontology lexica from language corpora by using association rules to identify correspondences between lexical elements on the one hand and ontological vocabulary elements on the other, in this paper we propose LexExMachinaQA, a framework allowing us to evaluate the impact of automatically induced lexicalizations in terms of alleviating the lexical gap in QA systems. Our framework combines the LexExMachina approach for lexicon induction with the QueGG system that relies on grammars automatically generated from ontology lexica to parse questions into SPARQL. We show that automatically induced lexica yield a decent performance i.t.o. F1 measure with respect to the QLAD-7 dataset, representing a 34% – 56% performance degradation with respect to a manually created lexicon. While these results show that the fully automatic creation of lexica for QA systems is not yet feasible, the method could certainly be used to bootstrap the creation of a lexicon in a semi-automatic manner, thus having the potential to significantly reduce the human effort involved.
    - **Venue:** LDK 2023

Future activities

Future work will extend this automatic creation of a lexicon for multilingual settings and enhance the performance in comparison to SOTA QALD systems.

Furthermore, the approach can handle a limited form of composition that allows the nesting of some rules into others. Future work will also investigate how to implement a more principled approach to semantic composition using Flexible Semantic Composition. The intention is to enable the community to contribute to and adapt the grammar generation to other languages, in particular for Indian and Arabic languages as the literature does not yet describe a QA system for an RDF dataset that works for those languages.

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.

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

Task Leaders:

- Leader: Simone Tedeschi

General Overview

The application of LLOD is primary in the context of WSD and EL, both in supervised and unsupervised approaches. In fact, an LLOD such as BabelNet (Navigli and Ponzetto, 2012) concurrently provides (i) a sound infrastructure to connect word senses across several languages by means of semantic relations (i.e. 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 contextualised 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 to further reduce the gap between English and other languages, as well as devise strategies to enhance the quality of the data they inherently feature.

Progress

- **Survey about WSD:**
  - **Title**: Recent Trends in Word Sense Disambiguation: A Survey
  - **Authors**: Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli
  - **Abstract**: Extensive overview of current advances in WSD, describing the state of the art in terms of i) resources for the task, i.e., sense inventories and reference datasets for training and testing, as well as ii) automatic disambiguation approaches, detailing their peculiarities, strengths and weaknesses.
  - **Venue**: ACL 2021
- **Entity Linking paper on overshadowed entities:**
  - **Title**: Focusing on Context is NICE: Improving Overshadowed Entity Disambiguation
  - **Authors**: Vera Provatorova, Simone Tedeschi, Svitlana Vakulenko, Roberto Navigli, Evangelos Kanoulas

    - **Abstract**: Entity overshadowing is a significant challenge for existing ED models: when presented with an ambiguous entity mentioned, the models are much more likely to rank a more frequent yet less contextually relevant entity at the top. Here, we present NICE, an iterative approach that uses entity-type information and graph-based relatedness to leverage context and avoid over-relying on the frequency-based prior.
    - **Venue**: Submitted to Natural Language Engineering journal (currently on ArXiv)
- **Coarse-grained WSD:**
    - **Title**: Analyzing Homonymy Disambiguation Capabilities of Pretrained Language Models
    - **Authors**: Lorenzo Proietti, Stefano Perrella, Simone Tedeschi, Giulia Vulpis, Leonardo Lavalle, Andrea Sanchietti, Andrea Ferrari and Roberto Navigli
    - **Abstract**: Traditional WSD systems rely on WordNet as the underlying sense inventory, often meticulously differentiating between subtle nuances of word meanings, which may lead to excessive complexity and reduced practicality of WSD systems in today's NLP. Here, we address these points and i) introduce a new large-scale resource that leverages homonymy relations to systematically cluster WordNet senses, and ii) we use this resource to investigate whether pretrained language models (PLMs) are inherently able to differentiate coarse-grained word senses.
    - **Venue**: LREC-COLING 2024
- **WSD for Latin:**
    - **Title**: Language Pivoting from Parallel Corpora for Word Sense Disambiguation of Historical Languages: a Case Study on Latin
    - **Authors**: Iacopo Ghinassi, Simone Tedeschi, Paola Marongiu, Roberto Navigli and Barbara McGillivray
    - **Abstract**: Recent studies have advanced the state-of-the-art in this task, but most of the work has been carried out on contemporary English or other modern languages, leaving challenges posed by low-resource languages and diachronic change open. In this work, we propose a new approach that exploits the WordNet structure as well as existing bilingual corpora to automatically produce training data for WSD in the Latin language.
    - **Venue**: LREC-COLING 2024

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 acknowledgments 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 favour 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

- **Term Relation Extraction**. Some efforts have been devoted to the relation extraction field, specifically, in the identification of relations between terms. In this line, several approaches have been tested:
    - **Title:** Thesaurus-enhanced annotation expansion to fine-tune a relation extraction model.
        - **Authors**: Patricia Martín Chozas, Artem Revenko
        - **Abstract**: In this paper we describe the design of an experiment to extract Hohfeld's deontic relations from legal texts. Our approach intends to minimise the manual effort in the annotation process by expanding a set of initial annotations with the legal domain knowledge contained in thesauri represented in Semantic Web formats. With such annotations, we perform a set of iterations to train a deep learning relation extraction model. After analysing the results, we will adapt the process to work on the extraction of Hohfeld's potestative relations. We also plan to use that model to recognise relations in unseen legal sub-domains.
        - **Venue**: DeepOntoNLP @ ESWC 2021
    - **Title**: Extraction and Semantic Representation of Domain-Specific Relations in Spanish Labour Law.
        - **Authors**: Artem Revenko, Patricia Martín Chozas

        - **Abstract**: Despite the freedom of information and the development of various open data repositories, the access to legal information to general audience remains hindered due to the difficulty of understanding and interpreting it. In this paper we aim at employing modern language models to extract the most important information from legal documents and structure this information in a knowledge graph. This knowledge graph can later be used to retrieve information and answer legal question. To evaluate the performance of different models we formalise the task as event extraction and manually annotate 133 instances. We evaluate two models: GRIT and Text2Event. The latter model achieves a better score of~ 0.8 F1 score for identifying legal classes and 0.5 F1 score for identifying roles in legal relations. We demonstrate how the produced legal knowledge graph could be exploited with 2 example use cases. Finally, we annotate the whole Workers' Statute using the fine-tuned Text2Event model and publish the results in an open repository.
        - **Venue**: SEPLN 2022
    - **Title**: Event Extraction and Semantic Representation from Spanish Workers' Statute Using Large Language Models
        - **Authors**: Gabriela Argüelles-Terrón, Patricia Martín-Chozas, Víctor Rodríguez-Doncel
        - **Abstract**: This work uses Large Language Models to process an important piece of Spanish legislation: the Workers' Statute. The proposed method extracts the relevant events in its articles using a GPT-3.5 model and represents the entities involved in the events and the relationships between them as RDF triples. The experiments carried out to select a high-performance strategy include both zero-and few-shot learning tests. Finally, this work proposes a strategy to uplift the extracted legal relations into a legal knowledge graph.
        - **Venue**: JURIX 2023
- **Terminology Modelling**. An important effort derived from this task is devoted to the representation of terminological resources in the Semantic Web (also called terminology modelling), that has been materialised as different actions:
    - Regular discussions are taking place pursuing an extension for the Ontolex model to the representation of terminologies, in the context of the W3C Ontology Lexica Community Group.
    - Joint organisation of three editions of the TermTrends event:
        - In 2022 as a tutorial within the EKAW conference, focused on discovering trends on terminology generation and modelling.
        - In 2023 as a workshop within the LDK conference, focused on terminology in the era of linguistic data science.
        - In 2024 (not yet celebrated) as a workshop within the MDTT conference, focused on models and best practices for terminology representation in the Semantic Web.
    - A paper derived from DFKI-UPM STSM in 2021:
        - **Title**: Representing terminological data in the Semantic Web
        - **Authors**: Patricia Martín-Chozas, Thierry Declerck, Elena Montiel-Ponsoda, Víctor Rodríguez Doncel
        - **Abstract**: This paper describes an approach to represent terminologies in the machine-readable format of the Semantic Web,

which improves the interoperability between terminological resources and opens up new possibilities yet to be discovered. The study's motivation stems from the realisation that the existing formalisms, such as SKOS or OntoLex-lemon, might not adequately capture the information within authoritative terminological resources. Therefore, we identified model requirements by formulating a set of Competency Questions derived from the analysis of terminological resources across various fields and domains, in line with the ontology development methodologies adopted in this work. During this analysis, we faced different representation challenges such as the various sources of term descriptions and the quality indicators related to terms. Consequently, we propose Termlex, a proposal based on the OntoLexlemon model that combines the conceptual structure of the SKOS model with the lexical information as modelled in OntoLex-lemon. In Termlex, we define new classes and properties to cover the specific needs of terminological resources coming from a variety of approaches. The paper concludes with the instantiation of the Termlex model through three different use cases that follow different modelling approaches as a validation attempt.

        - **Venue**: Terminology Journal 2024
    - A book chapter on the representation of terminologies containing metaphorical knowledge.
    - Efforts towards the conversion of existing terminologies and glossaries to Ontolex: the TERMCAT use case (paper to be submitted to SEPLN 2014).
- **Terminology Extraction**. Efforts were devoted to the creation of an evaluation framework for terminology extraction algorithms in Spanish. This framework consists, on the one hand, in a benchmark of five state-of-the-art term extraction algorithms, being two of them language model based and adapted to Spanish, and on the other hand, in two silver standards in Spanish translated from two of the main datasets used to evaluate this task in English: SemEval 2010 and SemEval2017. A paper with this contribution was sent to SEPLN 2024.
- **Definition Extraction**:
    - **Title**: Definitions in SNOMED CT through the lens of Terminology: from formal to textual
    - **Authors**: Sara Carvalho, Cornelia Wermuth and Rute Costa
    - **Abstract**: This paper aims to show how Terminology can help foster interoperability and more effective knowledge representation, organisation and sharing in the biomedical field, and on the other hand, support specialised communication among various stakeholders. SNOMED CT will be used to illustrate this, with the focus being on formal and textual (or natural language) definitions – the latter currently underrepresented in this resource - and on how a double dimensional terminological approach can benefit textual definition drafting, thereby assisting the work carried out by SNOMED CT national translation teams.
    - **Venue**: MDTT 2023

- Several invited talks on the generation, enrichment and publication of terminologies in the Semantic Web: EAFT[1], Translating and the Computer conference[2], ENDORSE follow-up events[3].

Future activities

In the framework of this task, discussions on the representation of terminologies in the Semantic Web will continue within the W3C Ontology Lexica Community Group working group. Also, thanks to the collaboration in this project, organisers of TermTrends expect to continue organising new editions of the event in the future. The focus of the workshop has been redirected to specifically tackle representation issues and standards for terminological resources, so it is aligned with the W3C discussions and the contributions to the workshop could be new input for the discussions.

[1] https://www.eaft-aet.net/en/summits/confirmed-speakers

[2] https://asling.org/tc45/

[3] https://op.europa.eu/en/web/endorse/follow-up-events

3. WG2 Survey Paper

The survey paper has been an ongoing effort in WG2 since the early days of NexusLinguarum, but didn't make any progress during the first half of the action. After modifications in the organisation of WG2, the survey was reborn and work restarted in a much more systematic and principled way.

Discussions led to the decision of not following a formal survey methodology to the letter (e.g. PRISMA), but to take inspiration from the general methodology. At first we initiated a systematic extraction of all identified venues for the community_based on subjective keywords, but the approach proved too time consuming and not reliable enough. We then reverted to a more classical approach, but with a twist!

We established the following work-plan:

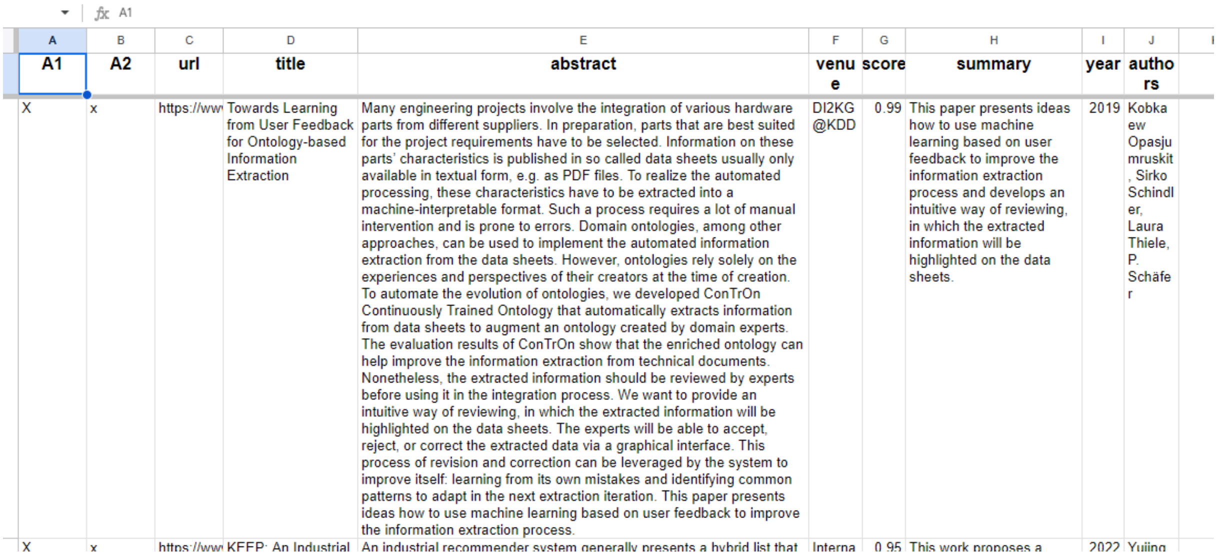
1. Definition of search queries for semantic scholar (aggregates most databases) based a a set of concepts/terms that
2. Filtering and automated pre-selection/scoring of the articles to favour a balanced number of concepts pertaining to LLD and to NLP
3. Manual screening phase based on titles and abstracts distributed across volunteers
4. Clustering to organise the papers more coherently and to help in reading assignments
5. Data extraction to retrieve the most pertinent information from the papers that could be expanded upon during writing
6. Writing

3.1. Screening guidelines

Each tab of a spreadsheet was assigned a main annotator and a reviewer. The workflow was the following:

1. The main annotator goes through the cluster and adds an **'X'** in the A column of the spreadsheet **to exclude a paper.** In most cases the annotator uses only the information in the spreadsheet, but when in doubt the link leads to the full semantic scholar page with more information and possibly the paper.
2. If the annotator is in doubt, they should highlight the line in yellow.
3. The reviewer checks the annotations of the main reviewer, adds a comment on lines where they do not agree, or gives additional feedback on yellow lines.

The process is iterative, even if the annotator hasn't finished, the reviewer can already start reviewing everything that's already annotated.



A	B	C	D	E	F	G	H	I	J
A1	**A2**	**url**	**title**	**abstract**	**venue**	**score**	**summary**	**year**	**authors**
X	x	https://ww	Towards Learning from User Feedback for Ontology-based Information Extraction	Many engineering projects involve the integration of various hardware parts from different suppliers. In preparation, parts that are best suited for the project requirements have to be selected. Information on these parts' characteristics is published in so called data sheets usually only available in textual form, e.g. as PDF files. To realize the automated processing, these characteristics have to be extracted into a machine-interpretable format. Such a process requires a lot of manual intervention and is prone to errors. Domain ontologies, among other approaches, can be used to implement the automated information extraction from the data sheets. However, ontologies rely solely on the experiences and perspectives of their creators at the time of creation. To automate the evolution of ontologies, we developed ConTrOn Continuously Trained Ontology that automatically extracts information from data sheets to augment an ontology created by domain experts. The evaluation results of ConTrOn show that the enriched ontology can help improve the information extraction from technical documents. Nonetheless, the extracted information should be reviewed by experts before using it in the integration process. We want to provide an intuitive way of reviewing, in which the extracted information will be highlighted on the data sheets. The experts will be able to accept, reject, or correct the extracted data via a graphical interface. This process of revision and correction can be leveraged by the system to improve itself: learning from its own mistakes and identifying common patterns to adapt in the next extraction iteration. This paper presents ideas how to use machine learning based on user feedback to improve the information extraction process.	DI2KG @KDD	0.99	This paper presents ideas how to use machine learning based on user feedback to improve the information extraction process and develops an intuitive way of reviewing, in which the extracted information will be highlighted on the data sheets.	2019	Kobkaew Opasjumruskit, Sirko Schindler, Laura Thiele, P. Schäfer
X	x	https://ww	KEEP: An Industrial	An industrial recommender system generally presents a hybrid list that	Interna	0.95	This work proposes a	2022	Yujing

Figure 1. Screening spreadsheet

NOTE: *for short, we will use NLP for NLP algorithm/systems/approaches and LD for LD/LLD/Semantic Web technologies*

We understand LD-aware NLP as:

- NLP that makes use of LD as an integral part of it
- NLP that is ready to use LD and is explicitly designed to do so to take advantage of LD capabilities such as dynamicity, interoperability, reasoning, etc.
- NLP that uses LD to describe or specify its pipelines, tasks, or processes to facilitate interoperability

We do NOT understand as LD-aware NLP:

- NLP that use or could use LD as a source of data merely, not being aware of the characteristics of LD
- NLP that is used to generate LD but without LD being relevant to the NLP algorithm(s)

Selection question:

Does this paper talk about LD aware NLP?

Examples:

YES:

- ○ An Information Extraction tool that uses a LD dataset as a source of entities and relations but also exploits the links to other entities.
- ○ An NLP system that uses an ontology to specify its pipeline/workflows.
- ○ A Machine Learning tool that exploits RDF bilingual dictionaries taking advantage of their links to infer new possible translations.

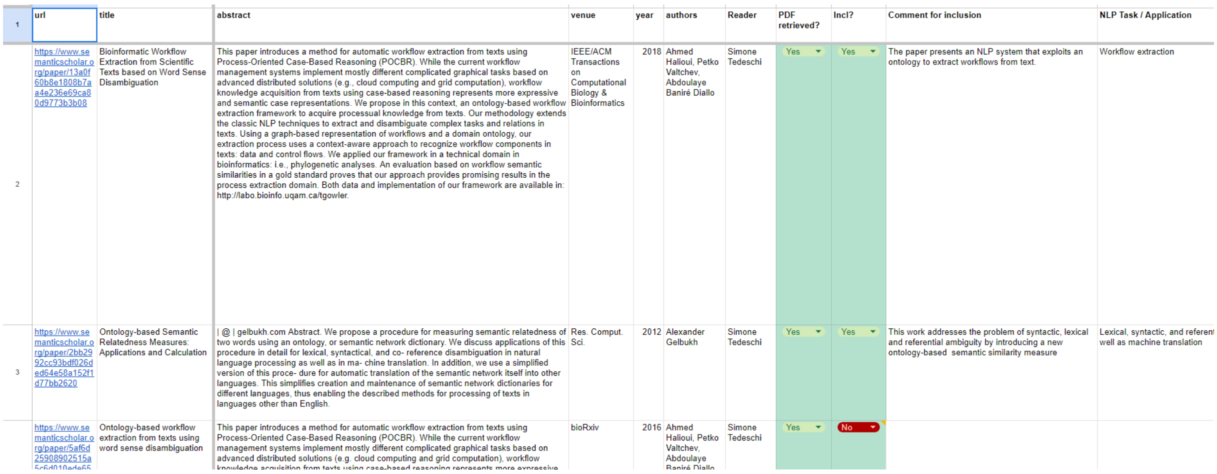
- A KG embedding method that uses Web-scale or cloud-scale federated learning on KGs, exploiting LD properties/capabilities (e.g. query federation) within the approach

NO:

- An Information Extraction tool that uses LD (e.g., DBpedia) only as a catalog of entities.
- An ontology learning/population tool that uses NLP techniques to extract semantic data from corpora in order to build/populate an ontology
- A LD resource (dataset, ontology,...) that could be potentially useful for NLP but without a proper evaluation, analysis, or concrete application, supporting this claim (every LD resource is potentially useful for NLP!).

3.2. Data Extraction

For the data extraction phase, the same guidelines were used, but this time people who were assigned reading tasks that involved reading the whole paper, double-checking inclusion if there was any doubt during screening, and extracting relevant information for the survey. Participants were asked to write a summary as if writing a paragraph of the survey, but also indicate the reasons for inclusion or exclusion, the NLP task concerned and extracting relevant cited references that weren't already included. Figure 2 illustrates the spreadsheet that was used to collect all the information.



1	url	title	abstract	venue	year	authors	Reader	PDF retrieved?	Incl?	Comment for inclusion	NLP Task / Application
2	https://www.semanticscholar.org/paper/13a0f60b8e1808b7aa4e236e69ca80d9773b3b08	Bioinformatic Workflow Extraction from Scientific Texts based on Word Sense Disambiguation	This paper introduces a method for automatic workflow extraction from texts using Process-Oriented Case-Based Reasoning (POCBR). While the current workflow management systems implement mostly different complicated graphical tasks based on advanced distributed solutions (e.g., cloud computing and grid computation), workflow knowledge acquisition from texts using case-based reasoning represents more expressive and semantic case representations. We propose in this context, an ontology-based workflow extraction framework to acquire processual knowledge from texts. Our methodology extends the classic NLP techniques to extract and disambiguate complex tasks and relations in texts. Using a graph-based representation of workflows and a domain ontology, our extraction process uses a context-aware approach to recognize workflow components in texts: data and control flows. We applied our framework in a technical domain in bioinformatics: i.e., phylogenetic analyses. An evaluation based on workflow semantic similarities in a gold standard proves that our approach provides promising results in the process extraction domain. Both data and implementation of our framework are available in: http://labo.bioinfo.uqam.ca/tgowler.	IEEE/ACM Transactions on Computational Biology & Bioinformatics	2018	Ahmed Halioui, Petko Valtchev, Abdoulaye Baniré Diallo	Simone Tedeschi	Yes ▾	Yes ▾	The paper presents an NLP system that exploits an ontology to extract workflows from text.	Workflow extraction
3	https://www.semanticscholar.org/paper/2bb2992cc93bdf026ded64e58a152f1d77bb2620	Ontology-based Semantic Relatedness Measures: Applications and Calculation	| @ | gelbukh.com Abstract. We propose a procedure for measuring semantic relatedness of two words using an ontology, or semantic network dictionary. We discuss applications of this procedure in detail for lexical, syntactical, and co- reference disambiguation in natural language processing as well as in ma- chine translation. In addition, we use a simplified version of this proce- dure for automatic translation of the semantic network itself into other languages. This simplifies creation and maintenance of semantic network dictionaries for different languages, thus enabling the described methods for processing of texts in languages other than English.	Res. Comput. Sci.	2012	Alexander Gelbukh	Simone Tedeschi	Yes ▾	Yes ▾	This work addresses the problem of syntactic, lexical and referential ambiguity by introducing a new ontology-based semantic similarity measure	Lexical, syntactic, and referent well as machine translation
	https://www.semanticscholar.org/paper/5af6d25908902515a5c6d010ede65	Ontology-based workflow extraction from texts using word sense disambiguation	This paper introduces a method for automatic workflow extraction from texts using Process-Oriented Case-Based Reasoning (POCBR). While the current workflow management systems implement mostly different complicated graphical tasks based on advanced distributed solutions (e.g. cloud computing and grid computation), workflow knowledge acquisition from texts using case-based reasoning represents more expressive	bioRxiv	2016	Ahmed Halioui, Petko Valtchev, Abdoulaye Baniré Diallo	Simone Tedeschi	Yes ▾	No ▾		

Figure 2. Data extraction spreadsheet. One tab per participant.

After this first phase of extraction is done, we will perform a secondary data extraction phase to process papers from secondary inclusion. Secondary inclusion covers all references extracted from papers read in the first phase as well as papers that were manually selected for addition through the expertise of the participants as well as recent publications in venues of interest.

3.3. Writing Phase

The introduction, and related literature review (i.e. regarding related surveys and survey methodologies) as well as the methods section (how the survey was conducted) will be written and consolidated by the WG2 co-leaders.

Then for the results section, once all relevant data is extracted, we will start the writing phase. All papers from the extraction table will be categorised according to a typology based on the extracted information. This will be the basis for the structure of the results section. Then, papers will be dispatched in relevant sub-sections and writers assigned by expertise to write the pertinent paragraphs.

A subset of the people who wrote for the results section will then be charged with writing the analysis and drawing the main highlights.

The WG2 co-leaders will then write the conclusions and perspectives as well as coordinate final screening and proofing efforts.

3.4. Tentative schedule for submission

The rate of progress with the current steps indicated a tentative end date for data extraction by the beginning of July 2024 and a start of writing efforts around September 2024. This would likely result in a first submission around the end of 2024.

4. Short Time Scientific Missions and Virtual Mobility Grants

The STSM (Short Term Scientific Mission) “Population of LLOD cloud with Deep learning approaches: Metaphor conceptualization and multilingual lexical relation acquisition” took place in September 2023, by Lucía Pitarch (University of Zaragoza) and Hugo Gonçalo Oliveira (University of Coimbra) at Zentrum für Translationswissenschaft (Austria), under the supervision of Dr. Dagmar Gromman.

The main objective of this STSM was contributing to task 4.3.1: BATS dataset translation. BATS stands for Bigger Analogy Test Set and was created by Gladkova et al, [1] to test analogical reasoning of Language Models. Current researchers have expanded its usage to other tasks such as probing the semantic knowledge encoded in Language Models. In nexus task UC.4.3.1. BATS was translated into over 15 languages providing a resource to explore cross-lingual knowledge in Language Models, which could then be used to populate ontologies. During the STSM several outcomes were produced: finalise the translation, validate the created dataset through relation acquisition and cross-lingual transfer experiments, and prepare the results for its submission to the LREC-COLING 2024 conference. Completing the mentioned tasks contributes to the following Nexus Linguarum tasks: 1.3 Cross-lingual data interlinking, access, and retrieval in LLOD (RDF representation of the dataset); 1.5 Development of the LLOD cloud for under-resourced languages and domains; 2.2 LLOD in Machine Translation (“translation-based analogy”); 3.3. Multilingual approaches, 2.1. Knowledge extraction, 3.2. Deep learning and 4.1.3. multilingual BATS dataset creation.

The future usage of the translated BATS dataset as well as its modelling as structured data, namely as RDF, was also discussed during the STSM, by doing so we aim at its linkage with other resources such as typological databases which might enable the study of other linguistic phenomena as lexical gaps or typological comparisons.

[1] Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t. In Proceedings of the NAACL Student Research Workshop, 8–15. 2016.

5. Organised Events

5.1. Summer Datathon on Linguistic Linked Open Data 2022

The 4th Summer Datathon on Linguistic Linked Open Data (SD-LLOD-22) was held physically from May 30th to June 3rd 2022 at Residencia Lucas Olazábal of Universidad Politécnica de Madrid, Cercedilla, Madrid (arrival expected on 29th evening).

The main goal of the SD-LLOD-22 datathon is giving people from industry and academia practical knowledge in the field of Linked Data applied to Linguistics. The final aim was to allow participants to migrate their own (or other's) linguistic data and publish them as Linked Data on the Web and/or develop applications on top of Linguistic Linked Data.

The the specific objectives of this datathon were:

- To generate and publish their own Linguistic Linked Data from some existing data sources.
- To apply Linked Data principles and Semantic Web technologies (Ontologies, RDF, Linked Data) into the field of language resources.
- To use the principal models used for representing Linguistic Linked Data, in particular OntoLex lemon.
- To learn about Linked Data-based NLP workflows and applications.
- To learn about potential benefits and applications of Linguistic Linked Data for specific use cases.

5.2. TermTrends22 Tutorial at EKAW

Trends in Terminology Generation and Modelling (TermTrends) tutorial was co-located within the 23rd International Conference on Knowledge Engineering and Knowledge Management (26-29 September 2022 - Bozen-Bolzano, Italy).

The focus of this tutorial was to study the different standardisation approaches, ranging from the initially proposed standards to represent terminology within ISO, to models that represent linguistic data in the Semantic Web, including emerging vocabularies still under development. Also, the tutorial presented an overview of the main features of the terminology work, exploring its evolution and new methods to speed up the terminology generation process. To complement this, different use cases were presented, in which both new methods to generate terminologies and new ways to represent terminological knowledge are applied in specific domains, such as Law or Life Sciences. Additionally, half of the tutorial was a hands-on session, testing several tools with different purposes, such as the extraction and enrichment of terminologies and their representation as per different vocabularies.

5.3. TermTrends23 Workshop at LDK

TermTrends 2023 was co-located with LDK 2023 at the University of Vienna, Austria, on September 13, 2023.

This half-day workshop provided a discussion forum regarding the theoretical and methodological approaches that have characterised Terminology in recent years, especially its central role in the organisation and sharing of specialised knowledge, both at a conceptual and linguistic level. In particular, we would like to focus on its connection to the Linguistic

Linked (Open) Data (LLOD) paradigm and to Semantic Web formats and technologies through the use of ontologies, thereby promoting the creation of interoperable terminological resources.

In addition, the workshop explored the advantages and challenges underlying various Terminology-related standardisation approaches, ranging from the initially proposed standards to represent terminology within the International Standardisation Organisation (ISO), to models that represent linguistic data in the Semantic Web, including emerging vocabularies still under development. Moreover, Terminology is currently benefiting from neural network approaches and architectures, and new methods are being proposed to enhance both term extraction and terminology enrichment tasks. In this line, the workshop also aimed to address novel ways to represent terminological knowledge, especially in multiple domains and applications, such as Digital Humanities, e-lexicography, or Life Sciences.

5.4. TermTrends24 Workshop at MDTT

TermTrends 2024, will be co-located with MDTT 2024 and aims to provide a discussion forum on the theoretical and methodological approaches for the representation of terminological data, both at a conceptual and a linguistic level. In particular, we would like to focus on their connection to the Linguistic Linked (Open) Data (LLOD) paradigm through the representation of these data according to Semantic Web formats. By adopting models or vocabularies proposed for the representation of linguistic data, we would contribute to the creation of interoperable and reusable terminological resources.

With this objective, the workshop intends to explore the advantages and challenges underlying various Terminology-related standardisation approaches, ranging from the initially proposed standards to represent terminology within the International Standardisation Organisation (ISO), such as the TermBase eXchange (TBX) format, to models that represent linguistic descriptions associated with ontologies in the Semantic Web, such as SKOS and Ontolex-lemon.

Being multidisciplinary in scope, it focuses on identifying terminological representation needs, as well as limitations of current models in addressing such needs, with the aim of also exploring the development of an extension of the Ontolex-lemon vocabulary and how that may contribute to overcoming such challenges.

6. Interaction with other WGs

6.1. WG1-WG2

This collaboration is based on efforts from members of the Action and from the Best Practices for Multilingual Linked Open Data (BPMLOD) W3C Community Group. The objective of this collaboration was to develop guidelines and best practices for linking data and services across languages. These guidelines discuss the creation of two sets of guidelines: one focusing on Linked Linguistic Open Data (LLOD), mainly pursued by Working Group 1, and the other on integrating LLOD with Natural Language Processing (NLP) services, pursued by Working Group 2. The guidelines are published in the form of a deliverable that highlights their contributions to the development of comprehensive recommendations for LLOD and LLOD-aware NLP services. Additionally, it discusses the evolution of the BPMLOD community, ongoing efforts to update guidelines, and plans for sustainability beyond the Nexus Linguarum COST Action.

6.2. WG2-WG3-WG4

In September 2023, Lucía Pitarch and Hugo Gonçalo Oliveira had a STSM in Vienna, with Dagmar Gromann. The STSM was focused on the compilation of the first version of MultiLexBATS, a dataset resulting from the translation of the lexico-semantic relations in BATS (Gladkova et. al, 2016) to 15 languages. This was followed by initial experiments with this dataset, where the main efforts included the evaluation of language models (BERT, XML-R, BLOOM) in the tasks of analogy completion and analogy-based translation. The work resulted in a paper later accepted for LREC-COLING (Gromann et. al, 2024). Other experiments were performed in the tasks of relation classification, visualisation, and representation of the dataset in an RDF format.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t. In Proceedings of NAACL 2016 Student Research Workshop, pages 8–15. ACL.

Gromann, D., Gonçalo Oliveira, H., Pitarch, L., Apostol, E.-S., Bernad, J., Bytyçi, E., Cantone, C., Carvalho, S., Frontini, F., Garabik, R., Gracia, J., Granata, L., Khan, F., Knez, T., Labropoulou, P., Liebeskind, C., di Buono, M. P., Anić, A. O., Rackevičienė, S., Rodrigues, R., Sérasset, G., Selmistraitis, L., Sidibé, M., Silvano, P., Spahiu, B., Sogutlu, E., Stanković, R., Truică, C.-O., Oleškevičienė, G. V., Zitnik, S., and Zdravkova, K. (2024). MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations. In Procs. LREC-COLING 2024 (to appear). ELRA and ICCL.

7. Future Directions

As NexusLinguarum comes to an end, we need to ensure that all that we have built within Nexus perdures and takes a life of its own, continuing in a sustainable manner even without the wonderful tools provided by COST. As described throughout this deliverable, the cornerstone of our sustainability strategy lies in establishing or reviving community groups and efforts that would pursue the efforts durably in time.

The BPMLOD W3C community group will sustain the creation of guidelines and their regular updates and evolutions, while the Ontolex-Lemon W3C community group will drive the development and adoption of community-wide standards by using BPMLOD recommendations as a springboard to bootstrap the development of said standards. While these communities can certainly continue functioning without any kind of financial support, we will strive to create a business plan that also ensures financial sustainability.

We submitted a COST Innovator Grant, with a focus on increasingly involving industrial actors and creating the foundation for long-term financing. Even if the CIG is not funded, many elements proposed therein could be supported through other community-led initiatives and by future project proposals that fund the development of specific aspects, much like the funding schemes used in many European infrastructures.

8. WG2 Related Publications

8.1. Publications Resulting from the Action

Ghinassi, Iacopo; Tedeschi, Simone; Marongiu, Paola; Navigli, Roberto; and McGillivray, Barbara (2024). Language Pivoting from Parallel Corpora for Word Sense Disambiguation of Historical Languages: a Case Study on Latin. Proceedings of LREC-COLING 2024.

Gromann, D., Gonçalo Oliveira, H., Pitarch, L., Apostol, E. S., Bernad, J., Bytyçi, E., Cantone, C., Carvalho, S., Frontini, F. … & Zdravkova, K. (2024). MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations. Proceedings of LREC-COLING 2024

Martín-Chozas, Patricia; Declerck, Thierry; Montiel-Ponsoda, Elena; and Rodríguez-Doncel, Víctor (2024). Representing terminological data in the Semantic Web: A proposal based on OntoLex-lemon. *Terminology* (online).

Gonçalo Oliveira, Hugo; Rodrigues, Ricardo; Ferreira, Bruno; Silvano, Purificação; and Carvalho, Sara (2024). BATS-PT: Assessing Portuguese Masked Language Models in Lexico-Semantic Analogy Solving and Relation Completion. PROPOR 2024.

8.2. Publications Related to the Action

Argüelles-Terrón, Gabriela; Martín-Chozas, Patricia; and Rodríguez-Doncel, Víctor (2023). Event Extraction and Semantic Representation from Spanish Workers' Statute Using Large Language Models. In Proceedings of the 36th International Conference on Legal Knowledge and Information Systems (JURIX 2023), Vol. 379, pp. 329-334.

Carvalho, Sara; Wermuth, Cornelia; and Costa, Rute (2023). Definitions in SNOMED CT through the lens of Terminology: from formal to textual. In: Di Nunzio, G., Costa, R. & Vezzani, F. (eds.), Proceedings of the 2nd International Conference on Multilingual digital terminology today. Design, representation formats and management systems (MDTT 2023), Lisbon, Portugal, June 29-30, 2023

Elahi, Mohammad Fazleh; Ell, Basil; and Cimiano, Philipp (2023). LexExMachinaQA: A framework for the automatic induction ofontology lexica for Question Answering over Linked Data. In Proceedings of the 4th Conference on Language, Data and Knowledge (pp. 207-218).

Elahi, Mohammad Fazleh; Ell, Basil; Nolano, Gennaro; and Cimiano, Philipp. Multilingual Question Answering over Linked Data building on a model of the lexicon-ontology interface. (Publication pending).

Martín-Chozas, Patricia and Revenko, Artem (2021). Thesaurus enhanced extraction of hohfeld's relations from spanish labour law. In Proceedings of the 2nd International Workshop on Deep Learning meets Ontologies and Natural Language Processing (DeepOntoNLP 2021) co-located with 18th Extended Semantic Web Conference, Vol. 2918, pp. 30-38.

Martín-Chozas, Patricia; Diab-Lozano, Isam; and Montiel-Ponsoda, Elena (2024). Representing metaphorical terms in the Semantic Web: A proposal from a sociocognitive perspective. Aspects of Cognitive Terminology Studies (Publication Pending).

Bevilacqua, Michele; Pasini, Tommaso, Raganato, Alessandro, & Navigli, Roberto (2021). Recent trends in word sense disambiguation: A survey. In International Joint Conference on Artificial Intelligence (pp. 4330-4338). International Joint Conference on Artificial Intelligence, Inc.

Gonçalo Oliveira, H. (2023). On the Acquisition of WordNet Relations in Portuguese from Pretrained Masked Language Models. In Proceedings of the 12th Global Wordnet Conference (pp. 41-49).

Gonçalo Oliveira, Hugo, and Rodrigues, Ricardo (2023). GPT3 as a Portuguese Lexical Knowledge Base?. In Proceedings of the 4th Conference on Language, Data and Knowledge (pp. 358-363).

Pitarch, Lucía; Bernad, Jorge; Dranca, Licri; Bobed Lisbona, Carlos.; and Gracia, Jorge (2023). No clues good clues: out of context Lexical Relation Classification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5607-5625).

Pitarch, Lucía; Bernad, Jorge; and Gracia, Jorge (2023). MEAN: Metaphoric Erroneous ANalogies dataset for PTLMs metaphor knowledge probing. In Proceedings of the 4th Conference on Language, Data and Knowledge (pp. 147-152).

Pitarch, Lucía; Bobed, Carlos; Avián, David; Gracia, Jorge; and Bernad, Jorge (2024). Building MUSCLE, a Dataset for MUltilingual Semantic Classification of Links between Entities. Proceedings of LREC-COLING 2024.

Poikkimäki, H., Leskinen, P., Tamper, M., and Hyvönen, E. (2022). Analyses of networks of politicians based on linked data: Case ParliamentSampo–Parliament of Finland on the Semantic Web. In European Conference on Advances in Databases and Information Systems (pp. 585-592). Cham: Springer International Publishing.

Proietti, Lorenzo; Perrella, Stefano; Tedeschi, Simone; Vulpis, Giulia; Lavalle, Leonardo; Sanchietti, Andrea; Ferrari, Andrea; and Navigli, Roberto (2024). Analyzing Homonymy Disambiguation Capabilities of Pretrained Language Models. Proceedings of LREC-COLING 2024.

Provatorova, Vera; Tedeschi, Simone; Vakulenko, Svitlana; Navigli, Roberto, and Kanoulas, Evangelos (2022). Focusing on Context is NICE: Improving Overshadowed Entity Disambiguation. arXiv preprint arXiv:2210.06164.

Revenko, Artem and Martín-Chozas, Patricia (2022). Extraction and Semantic Representation of Domain-Specific Relations in Spanish Labour Law. Procesamiento del Lenguaje Natural, 69, 105-116.

Zdravkova, Katerina (2022). Resolving Inflectional Ambiguity of Macedonian Adjectives. Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 60–67