# NLP

#### Time: Thursday, September 21, 2023 - 16:45 to 18:00

#### Chair: TBA

## **Talks**

### Perplexed by idioms?

Purpose: In an information theoretic framework, this study aims to identify idiomatic expressions (IE) in English by measuring perplexity between texts with IE, without IE (literals) and large reference texts.

Methodology: Four texts are analysed: one text with at least one IE per sentence, one text without any IE, and two large, sentence-wise unrelated reference texts . Perplexity is measured based on bi- till heptagrams of PoS tags and tokens, and , in addition, of thematic roles within the boundaries of a sentence extracted by the LOME parser. We assume that the divergence of perplexity from the reference would be higher for IEs compared to literals.

Findings: We observed that there is no significant difference in perplexity between IE, literal expressions and reference texts.

Value: Model-based identification of IE can be useful in automatic speech processing applications, which sometimes have problems with IE. We suggest utilising larger extra-sentential contexts to determine perplexity. Additionally, reducing the number of thematic roles can avoid an uniform distribution of n-grams.

| Michael Richter | Michael Richter holds a Dr. degree (2000) from Katholieke Universiteit Nijmegen. He completed his habilitation at Justus-Liebig-Universität Gießen in 2023. Richter’s recent research focus is directed on communication models in natural language and information theory.  He has served as a researcher at Leipzig University and InfAI since 2017. Richter was a Visiting Professor at Justus-Liebig-Universität Gießen, holding a C4-Professorship in Applied Linguistics and Computational Linguistics. Prior to this, he worked as a lecturer and postdoc at academic institutions including Rijksuniversiteit Groningen, Radboud Universiteit Nijmegen, and others. |
| --- | --- |

### Why you might not be ready for LLM's [SP]

A key shortcoming of LLM's is they require a high quality Knowledge Graph - Have you been -

* Seduced by the hype - (or worse management have been seduced by the hype)
* Have no formal or informal catalogue of internal data assets
* Have no common glossary of terms or industry standards
* Have limited access to SMEs or their documentation
* Are not in a data oriented green fields project

| John PlacekCo-founder of Semantic Partners[Affiliation page](https://www.semanticpartners.com/) | A Software and Semantic Engineer with 15+ years experience collaborating with clients to deliver value through iterative development, deliberate discovery and validated learning. Experienced in leading and coaching delivery teams in government, finance and transport. One of the founders of Semantic Partners, John is currently working on a key client in the Investment Banking space delivering a data mesh. |
| --- | --- |

### QALD-9-ES: A Spanish Dataset for Question Answering Systems [Online]

Knowledge Graph Question Answering (KGQA) systems enable access to semantic information for any user who can compose a question in natural language. KGQA systems are now a core component of many industrial applications, including chatbots and conversational search applications. Although distinct worldwide cultures speak different languages, the number of languages covered by KGQA systems and its resources is mainly limited to English. To implement KGQA systems worldwide, we need to expand the current KGQA resources to languages other than English. Taking into account the recent popularity that Large-Scale Language Models are receiving, we believe that providing quality resources is key to the development of future pipelines. One of these resources is the datasets used to train and test KGQA systems. Among the few multilingual KGQA datasets available, only one covers Spanish, i.e., QALD-9. We reviewed the Spanish translations in the QALD-9 dataset and confirmed several issues that may affect the KGQA system's quality. Taking this into account, we created new Spanish translations for this dataset and reviewed them manually with the help of native speakers. This dataset provides newly created, high-quality translations for QALD-9; we call this extension QALD-9-ES. We merged these translations into the QALD-9-plus dataset, which provides trustworthy native translations for QALD-9 in nine languages, intending to create one complete source of high-quality translations. We compared the new translations with the QALD-9 original ones using language-agnostic quantitative text analysis measures and found improvements in the results of the new translations. Finally, we compared both translations using the GERBIL QA benchmark framework using a KGQA system that supports Spanish. Although the question-answering scores only improved slightly, we believe that improving the quality of the existing translations will result in better KGQA systems and therefore increase the applicability of KGQA w.r.t. the Spanish language domain.

| Javier Soruco Lopez |
| --- |

### 