UNIVERSITÀ DEGLI STUDI DI SALERNO

THE LINGUISTIC LINKED OPEN DATA THROUGH THE LINGUISTS' LENS

Data Quality meets Machine Learning ad Knowledge Graphs

Pasquale Esposito, PhD student in Linguistics - 26 May 2024

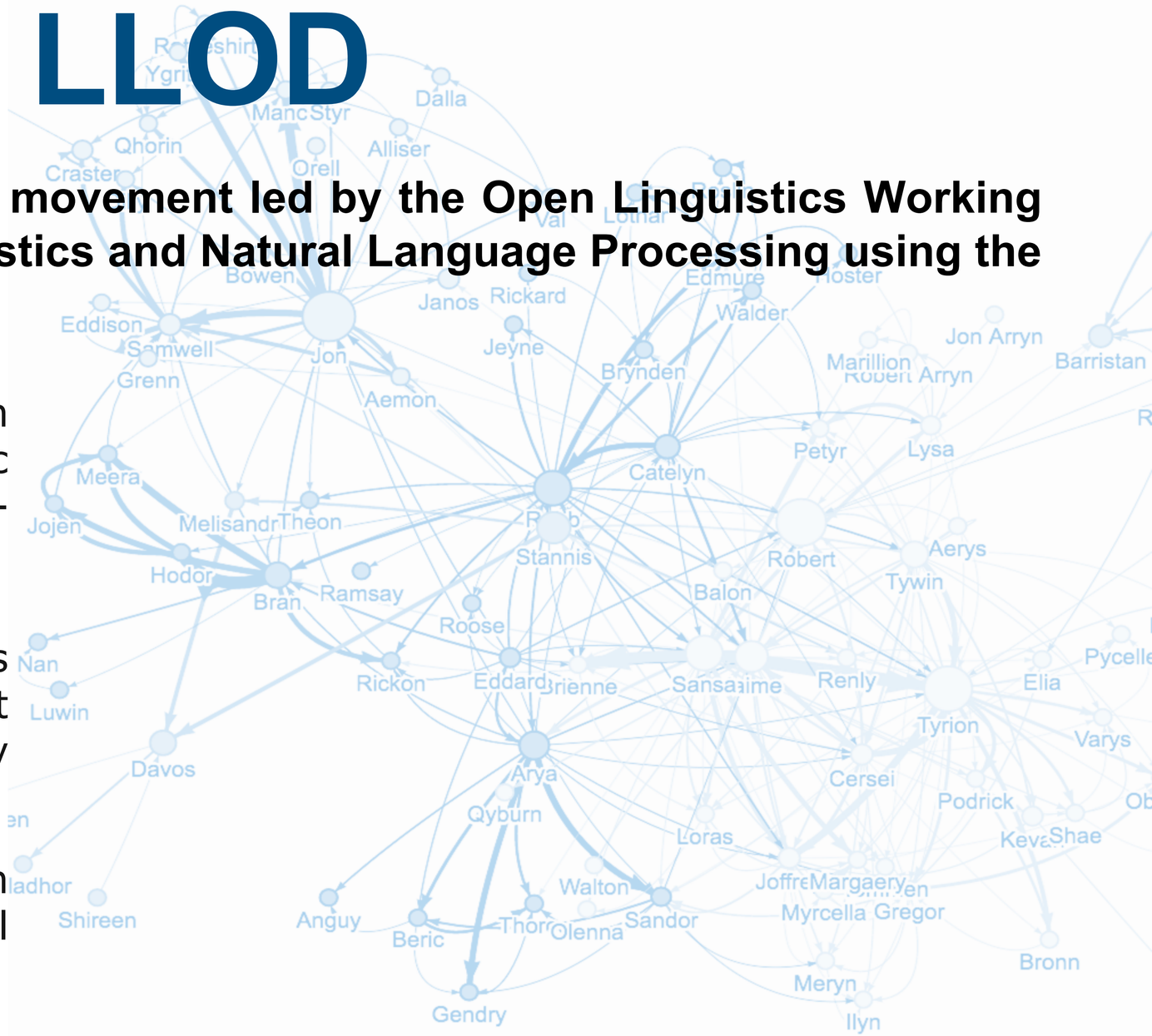
Introduction to LLOD

Linguistic Linked Open Data (LLOD) is a movement led by the Open Linguistics Working Group and aims to publish data for linguistics and Natural Language Processing using the linked data (LD) principles.

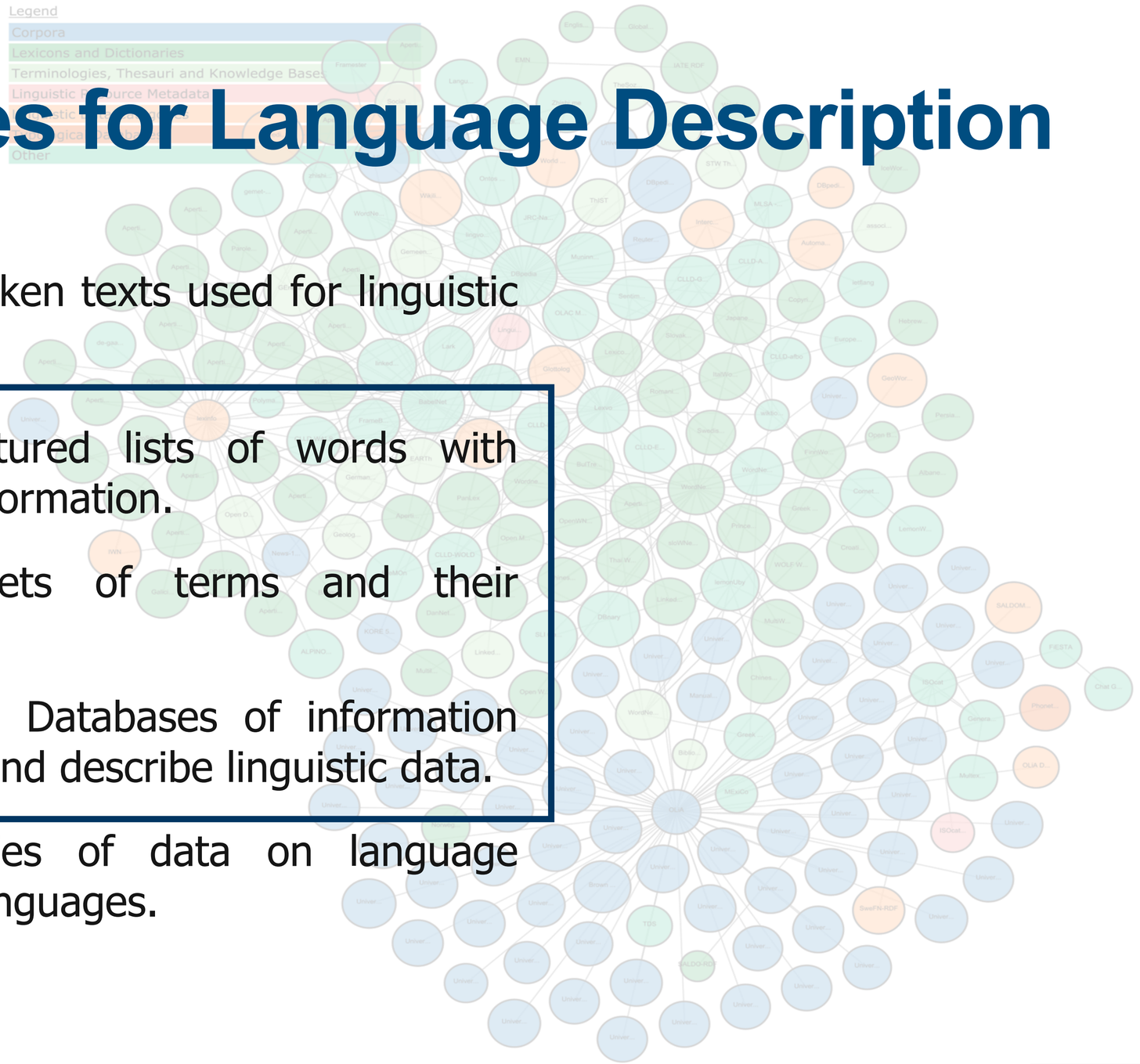
- **Linguistic Linked Open Data (LLOD)** is an initiative to create a web of interlinked linguistic resources that are openly accessible and machine-readable.

- **Importance in Research:** LLOD supports linguistic research by providing structured data that facilitates metadata enrichment dictionary generation and more.

- **Relevance to AI:** LLOD plays a crucial role in enhancing AI applications particularly in Natural Language Processing (NLP) tasks.



Linguistic Resources for Language Description



Corpora: Collections of written or spoken texts used for linguistic research.

Lexicons and Dictionaries: Structured lists of words with meanings usage and other linguistic information.

Terminologies and Thesauri: Sets of terms and their relationships within specific domains.

Knowledge Bases and Metadata: Databases of information and data categories used to structure and describe linguistic data.

Typological Databases: Repositories of data on language typologies and characteristics across languages.

The Linguistic Linked Open Data Cloud from lod-cloud.net

Efforts to support linguistic side of data

- **Convenient systems and directives for linguistic research** with efforth which have been focused on creating systems and directives that facilitate linguistic research

- **Dictionary Generation:** Creating dictionaries from encyclopedic knowledge to support linguistic research

- **Metadata enrichment** from encyclopedic knowledge

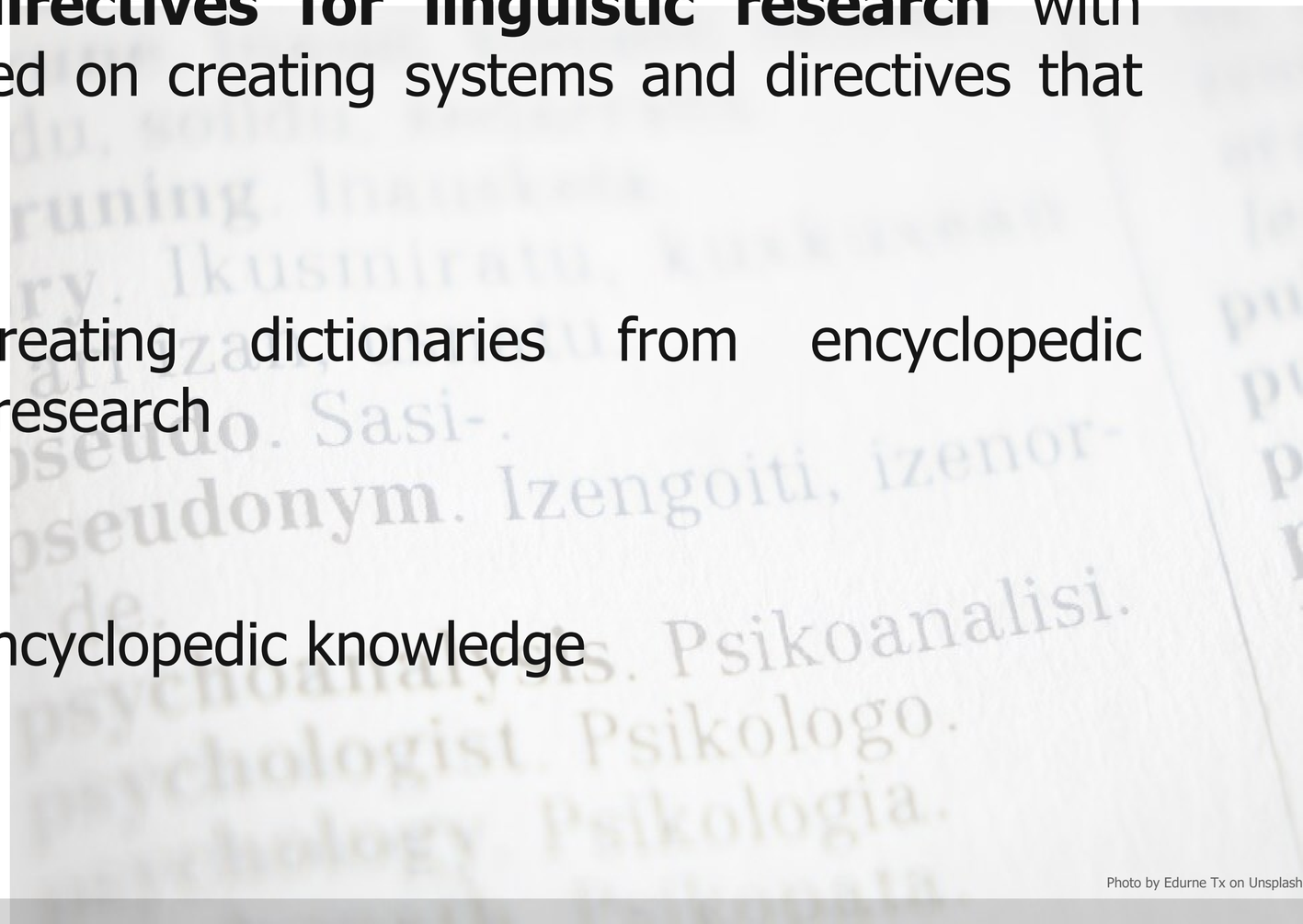


Photo by Edurne Tx on Unsplash

Machine-Readable Data and Applications in AI

- **Large Language Models (LLMs):** Utilizing machine-readable data to enhance LLMs and word embeddings.
- **Word-Sense Disambiguation:** Improving the accuracy of word-sense disambiguation in NLP tasks.
- **AI and NLP Applications:** Applications include meaning representation personal knowledge graph representation and cross-language linking.
- **Machine-readable and exploitable data:** Data is kept machine-readable for tasks such as LLMs improvement word embeddings and more.

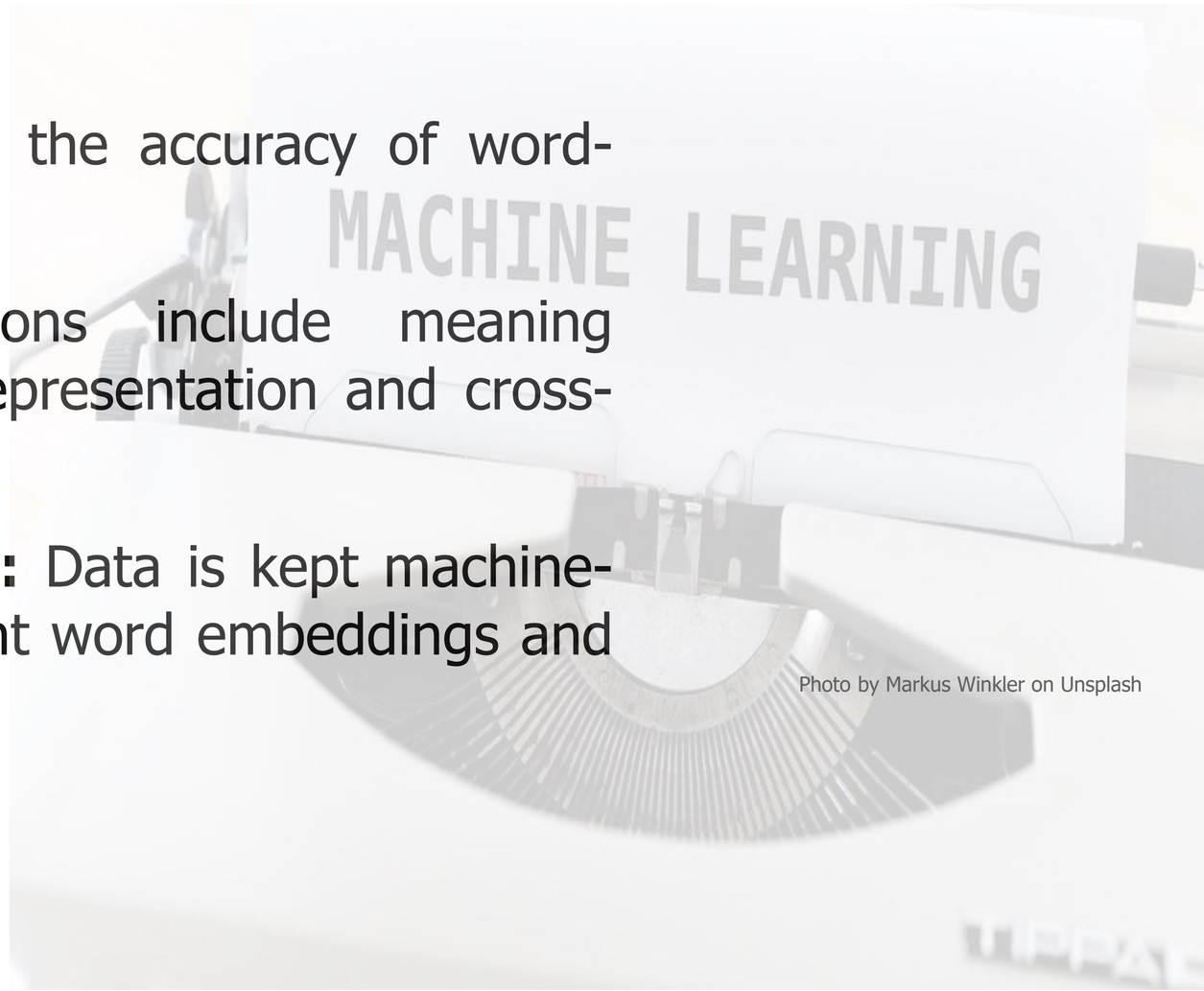


Photo by Markus Winkler on Unsplash

Do LLOD represent (all) expected features of a traditional lexicographic resource?

Linguists’ expectations from LLOD’s lexicons

Entry
Form
Definition
Senses
Phonetic
Transcription
Morphological pattern
Domain label
Different usages
Register label
Style label

Relevance
Animacy
Aspect
Case
Clitic
Definiteness degree
Finiteness
Gender
Number
Modification type
Part of speech
Person
Tense

Introduction to OntoLex Lemon

- **OntoLex Lemon (or simply OntoLex)** is a model designed for representing **lexical information** in the Semantic Web and within Linked Data frameworks.

- It is especially tailored to handle multilingual and semantic lexical data effectively. The model is developed and maintained by the W3C Ontology-Lexica Community Group.

OntoLex-Lemon

CORE MODULE

LexicalEntry: Represents individual words or multi-word expressions.

Lexicon: A container for a set of LexicalEntry objects, typically defined for a specific language or domain.

- **Semantic Linking**:

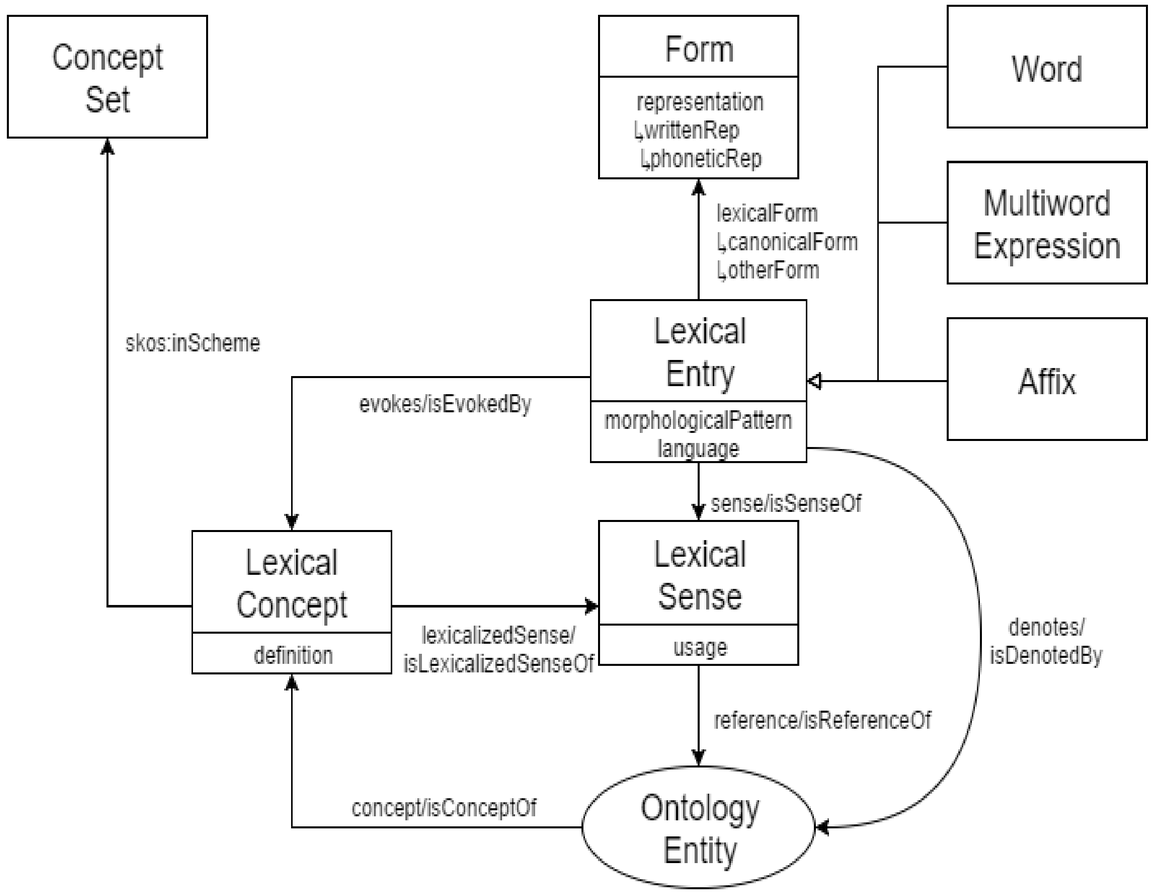
LexicalSense: Connects LexicalEntry objects to their meanings, linking them to concepts defined in ontologies, thus facilitating a deeper semantic integration.

- **Morphological and Syntactic Representation**:

Addresses the representation of morphological forms and syntactic behaviors of lexical entries, allowing for a comprehensive description of language elements.

- **Multilinguality**:

Supports the representation of translations and multilingual lexicalizations, making it suitable for applications like multilingual dictionaries and semantic translation tools.



https://www.w3.org/2016/05/ontolex/

Additional Modules in OntoLex Lemon

OntoLex-Lemon: Core module

OntoLex-SynSem: for Syntax and Semantics

OntoLex-Decomp: for Decomposition

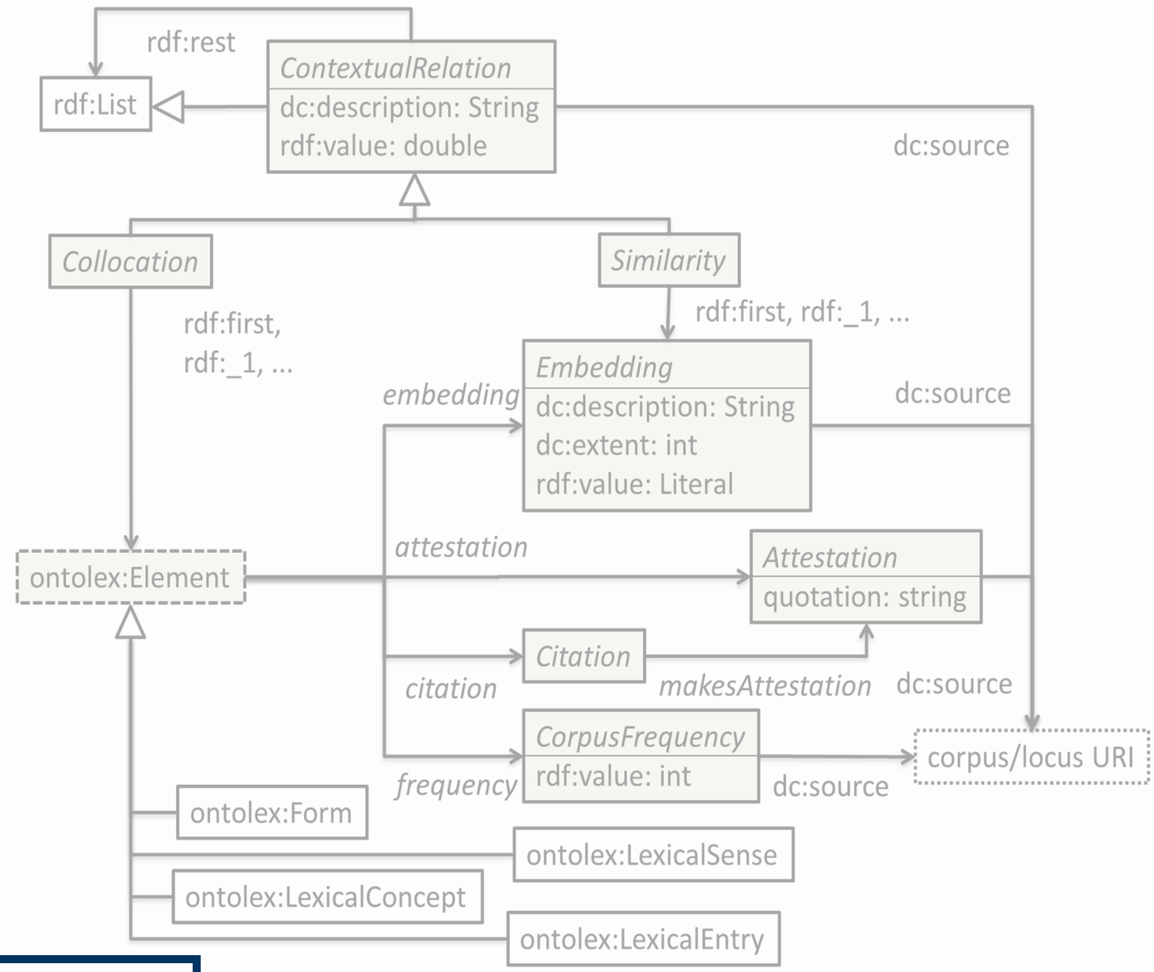
OntoLex-VarTrans: for Variation and Translation

OntoLex-LiMe: for Linguistic Metadata

OntoLex-Lexicog: module for Lexicography

OntoLex-Morph: for Morphology

OntoLex-FrAC: emerging OntoLex module for Frequency, Attestation and Corpus-Based Information



Linguists' expecations from LLOD's lexicons

- ✓ Entry
- ✓ Form
- ✓ Definition
- ✓ Senses
- ✓ Phonetic transcription
- ✓ Morphological pattern
- ✗ Domain label
- ✓ Different usages
- ✗ Register label
- ✗ Style label

- ✗ Relevance
- ✗ Animacy
- ✓ Aspect
- ✓ Case
- ✗ Clitic
- ✗ Definiteness degree
- ✗ Finiteness
- ✓ Gender
- ✓ Number
- ✓ Modification type
- ✓ Part of speech
- ✓ Person
- ✓ Tense

Formality in speech

In linguistics, "formality" refers to the degree to which language, vocabulary, and expressions conform to established or conventional standards that are typically associated with more serious, professional, or polite contexts.

Key Aspects of Formality in Linguistics:

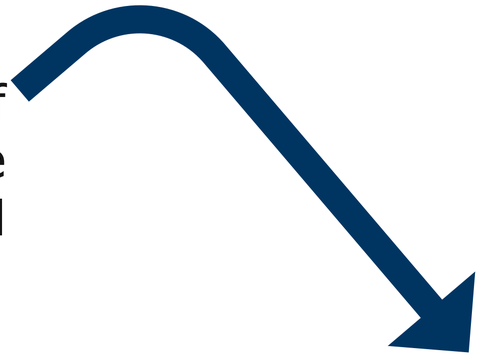
- Vocabulary Selection
- Sentence Structure
- Tone and Politeness
- Purpose and Function
- Register

How missing features can be represented?

Annotations: Register can be represented using annotations. Lexical entries can be annotated with specific properties that describe the register, such as "formal," "informal," "technical," or "colloquial.»

Custom Extensions: Given the modular nature of OntoLex, developers can create custom extensions to the model to explicitly handle register as a distinct lexical feature

Use of Existing Modules: existing modules like the Lexico-Semantic Module or the Variants Module might be adapted to include aspects of register.



Perception of formality is psychologically determined and cannot be considered a static picture as the other lexicographic features. Reporting formality for computational aims cannot be limited to single labels.

Calculating text-formality

First attempt to apply a computational measure to large corpora of linguistic data, without requiring specific rules for handling all possible subtleties or exceptions of the particular language or situation.

The formula should be capable to unambiguously distinguish discourses that are considered formal from those that are considered informal.

$$F\,[\textit{formality measure}] = \frac{\textit{noun frequency} + \textit{adjective frequency} + \textit{preposition frequency} + \textit{article frequency} - \textit{pronoun frequency} - \textit{verb frequency} - \textit{adverb frequency} - \textit{interjection frequency} + 100}{2}$$

(Heylighen and Dewaele, 1999)

Computational scheme formality weights

		"formal" categories			*"deictic" categories*				
	Nouns	**Articles**	**Prepos.**	**Adject.**	**Pronouns**	**Verbs**	**Adverbs**	**Conjun.**	**Formality**
Oral Female	10.40	6.89	5.86	8.09	16.95	19.35	17.45	7.47	**38.7**
Oral N.Acad.	12.75	8.50	6.34	6.71	16.01	18.80	19.31	6.34	**40.1**
Oral Male	11.48	8.16	6.69	7.63	15.84	18.45	16.53	7.05	**41.6**
Oral Acad.	13.16	9.58	7.91	7.13	13.96	17.75	17.88	7.13	**44.1**
Novels	18.52	10.48	10.26	10.00	13.25	20.62	10.47	6.06	**52.5**
Fam. Magaz.	21.78	9.77	12.21	11.14	10.09	18.71	9.74	6.39	**58.2**
Magazines	24.20	11.61	13.90	10.93	8.55	17.68	8.73	4.34	**62.8**
Scientific	23.10	15.00	13.75	10.75	6.71	16.58	7.98	5.98	**65.7**
Newspapers	25.97	14.68	14.54	10.57	5.62	16.69	7.21	4.70	**68.1**

(Heylighen and Dewaele, 1999)

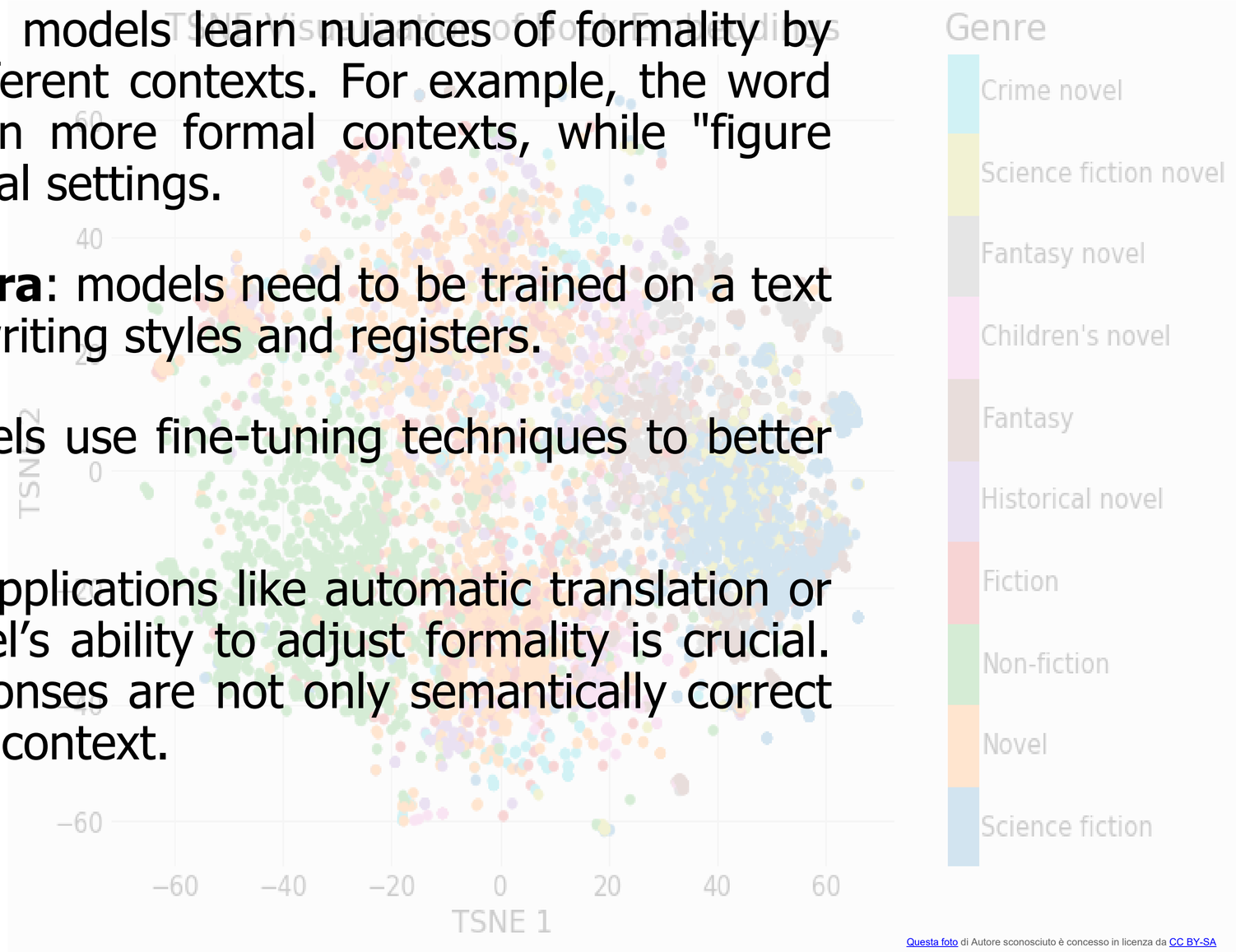
How Formality is managed in Distributional Semantics Models:

Learning from Variable Contexts: models learn nuances of formality by observing how words are used in different contexts. For example, the word "calculate" might frequently appear in more formal contexts, while "figure out" might be more common in informal settings.

Training on Heterogeneous Corpora: models need to be trained on a text corpus that includes a wide range of writing styles and registers.

Model Adaptability: Advanced models use fine-tuning techniques to better adapt to specific levels of formality.

Use in Practical Applications: In applications like automatic translation or automated response systems, a model's ability to adjust formality is crucial. This ensures that the generated responses are not only semantically correct but also stylistically appropriate to the context.



Register Labels in Traditional OR Synset-based resources

Synset based resources:

- Synset is monolingual and aggregate synonyms in a single “node”
- Words are denotationally equivalent and can be substituted for one another in many, but not all, contexts

Results:

- Formality relevance measures are absent also for lexical units in the synset-based.
- We detect sensitivity towards regional variation in OEW.

WordNet
Open English Wordnet

Non-synset resources:

•Wiktionary embraces and explicitly reports the linguistic labels for domain and style and register, providing the related and expected information for the lexeme.
•DBnary is the ontology-based representation of Wiktionary modeled according to a modified version of the OntoLex model.
•Conversely from Wiktionary, DBnary does not report registers or domain labels.

Wiktionary
DBnary

Non synset and machine learning/embeddings dependent

- ConceptNet is generated on an embedding-based structure a hybrid framework between distributional semantics and relational knowledge
- lexical description does not foresee any formality relevance measure.

ConceptNet

Benefits for AI

- The Semantic Web community might take advantage of manually annotated and linguistic-validated corpora to include formality weights and model it as valuable data

- The LLMs could take advantage of pure natural language processing operations and implement the linguistics side with data management and AI principles in order to offer systems that can interact and automatically mirror formality in language in a human-like way

Conclusions

- The common ground between linguistics, Natural Language Processing, Semantic Web, and AI could result in the usage of corpora as a qualitative and quantitative reference to analyze and then compute the authentic reproduction of speech

- A weighted and precise description of formality relevance for LLOD satisfies **linguistic expectations** and produce a wide range of benefits in several (semi-) automatic AI-driven applications

- We suggest a hybrid corpus-based/crowd-sourced approach to detect formality weights for lexicons of different languages as a starting point towards the integration of computational formality measures to linguistics resources of the LLOD cloud

UNIVERSITÀ DEGLI STUDI DI SALERNO

Thank you very much for your attention!

pasesposito@unisa.it

References