A Multilingual Linguistic Domain Ontology

Mariem Neji
MIRACL Laboratory
Sfax, Tunisia
mariem.neji@yahoo.fr

Fatma Ghorbel
CEDRIC Laboratory
Paris, France
fatmaghorbel@gmail.com

Bilel Gragouri
MIRACL Laboratory
Sfax, Tunisia
bilel.gargouri@fsegs.rnu.tn

Nada Mimouni
CEDRIC Laboratory
Paris, France
nada.mimouni@lecnam.net

Elisabeth Métais
CEDRIC Laboratory
Paris, France
elisabeth.metais@cnam.fr

Abstract

Natural Language Processing provides a very significant contribution to various research areas such as the e-health, e-business, education and antiterrorism. However, understanding the meaning, scope and usage of linguistic knowledge is a tedious task for a heterogeneity of potential users. Several approaches have been proposed to represent heterogeneous linguistic knowledge covering specific features of some languages. However, these approaches focused only on the data aspect of linguistic knowledge and neglected the processing one. Moreover, most of them do not support multilingualism and lack of powerful semantic representation and reasoning abilities. In this paper, we propose a multilingual linguistic domain ontology, called LingOnto, that represents and reasons about *(1)* linguistic data, *(2)* linguistic processing functionalities and *(3)* linguistic processing features. Our ontology supports English, French and Arabic languages and can be used by linguistically under-skilled users. In order to evaluate LingOnto and measure its efficiency, we applied it to a framework of identifying valid composition workflows of linguistic web services. Finally, we give the results of the carried-out experiments.

1 Introduction

Natural **L**anguage **P**rocessing (NLP) provides a very significant contribution to various research areas such as the e-health, e-business, education and antiterrorism. It has witnessed over the last years an acceleration in progress on a wide range of different applications such as sentiment analysis, knowledge mining and reasoning and search engine (Zhou et al., 2020).

However, understanding the meaning, scope and usage of linguistic knowledge is a tedious task. This complexity is mainly due to three reasons. First, NLP domain's potential users are heterogeneous and a considerable number of them are linguistically under-skilled. Second, the language is always changing, evolving, and adapting to its user's needs. For instance, words can acquire new meanings over time (e.g., the meaning of **"apple"** is a fruit but the meaning of **"Apple"** is a company). Finally, every language has its own specificities. Indeed, each language has its own structures and ways of interpreting. For example, Arabic has verbal and nominal sentences; but English has only verbal sentences.

In NLP, two types of approaches have been proposed to represent linguistic knowledge : *(1)* online registries-based approach such as the ISOcat registry[1], the SIL Glossary of linguistic terms[2] and the CLARIN Concept Registry (Schuurman et al., 2016) and *(2)* ontologies-based approach such as the General Ontology for Linguistic Description (GOLD) (Farrar and Langendoen, 2010), OntoTag ontologies (De Cea et al., 2004) and WordNet (Gangemi et al., 2002). However, these approaches present only linguistic data (e.g., word, noun, verb and adjective) and do not focus on modeling linguistic processing functionalities (e.g., tokenization, stemming and part of speech tagging) and linguistic processing features (e.g., treatment type, formalism and process-

[1] http://www.isocat.org/

[2] http://www-01.sil.org/linguistics/GlossaryOflinguisticTerms

ing level). Moreover, most of these approaches lack of powerful semantic representation and reasoning abilities. Finally, most of them do not support multilingualism.

In this paper, we propose a multilingual linguistic domain ontology, called LingOnto. It represents and reasons about linguistic processing functionalities, features and their linking with linguistic data rather than merely representing linguistic data. It supports English, French and Arabic languages. This ontology can be used by linguistically under-skilled users. In order to evaluate LingOnto, we choose to experiment it in the context of lingware engineering (Baklouti et al., 2010). Particularly, it is applied to a framework of identifying valid composition workflows of **Ling**uistic **W**eb **S**ervices (LingWS).

The current paper is organized as follows. Section 2 presents the related work. In Section 3, we detail the proposed ontology. Section 4 describes the carried-out experiments and the obtained results. Finally, Section 5 draws conclusions and future research directions.

2 Related Work

A considerable number of approaches for representing linguistic knowledge are available in the literature. We categorize them into two categories ***"online registers-based approach"*** and ***"ontologies-based approach"***.

2.1 Online Registers-Based Linguistic Knowledge Representation Approach

The SIL Glossary of linguistic terms (Eugene et al., 2004) provides information in the form of glossaries and bibliographies designed to support linguistic research. However, only 900 linguistic terms are covered in this glossary. Moreover, this latter supports only English and French languages. In addition, the usage of the SIL glossary is limited to search a defined term whose relation to other terms is unspecified. Consequently, it is not suitable for gaining comprehensive knowledge about a linguistic term in the NLP field.

In an attempt to provide a more comprehensive registry, (Kemps-Snijders et al., 2009) proposed ISOcat Data Category Registry (DCR). This registry aims at representing data categories at different linguistic levels such as syntactic, morphosyntactic, terminological, lexical and so on. However, ISOcat provides a wide range of different "views" and "groups" which makes navigating through it a very hard task. Moreover, it has no data model representing linguistic terminology in an interrelating holistic structure. Besides, its semantic structure provides definitions and very unspecific superordinate and subordinate concept relations such as "is_a" or "has_kinds".

Trying to define linguistic data in a stricter manner, the CLARIN Concept Registry, has taken over the work of ISOcat. Although, it still provides very limited structural and relational information.

2.2 Ontologies-Based Linguistic Knowledge Representation Approaches

(Farrar and Langendoen, 2010) proposed the GOLD ontology. It is based on the principles of knowledge engineering. It provides a taxonomy of nearly 600 linguistic concepts and formalizes 83 objects properties. These latter are very complex, specific and interrelate mostly only two concepts, which leaves the majority of the concepts unrelated. In addition, GOLD originates from the language documentation community and do not focus on NLP and corpus interoperability. Therefore, a number of data categories commonly assumed in NLP were not originally represented in GOLD. For example, gold:CommonNoun was added only recently following a suggestion by the author. Moreover, it do not cover all the linguistic knowledge. It defines only linguistic data and do not focus on modelling linguistic processing functionalities and features. A more fundamental problem is that this ontology is a very inefficient model for linguistic terminology. GOLD conflate both semantic and syntactic roles. The development of GOLD process has been stopped in 2010.

Focusing on the interoperability and language understanding, (De Cea et al., 2004) proposed OntoTag ontologies. These ontologies are applied to develop NLP applications on the basis of ontological representations of linguistic annotations. However, they consider only Iberian Romance languages (in particular Spanish). Moreover, they cover only linguistic data. They are not publicly available at the moment.

(Chiarcos and Sukhreva, 2015) proposed OLiA

ontologies which are closer related to the Onto-Tag ontologies. They introduce an intermediate level of representation between ISOcat, GOLD and other repositories of linguistic reference terminology. However, these ontologies do not represent and reason about linguistic processing and features.

WordNet is a lexical resource that is rich enough to be considered alongside actual ontologies. It contains an extensive taxonomic and mereological structure which could be regarded as a kind of proto-ontology. However, it focuses on representing only some linguistic data. Moreover, WordNet object properties are not used in a consistent way, sometimes they are broken or present redundancy. (Gangemi et al., 2002) demonstrated that a substantial transform of WordNet's upper categories is needed in order to be used directly as an ontology.

It is worth mentioning that, in all of the above mentioned published works, the authors focused only on the data aspect of linguistic knowledge and neglected the processing one. Moreover, most of them do not support multilingualism and lack of powerful semantic representation and reasoning abilities.

3 LingOnto: a Multilingual Linguistic Domain Ontology

We propose a multilingual linguistic domain ontology, called LingOnto. It represents and reasons about linguistic knowledge. It handles *(1)* linguistic data *(2)* linguistic processing functionalities and *(3)* linguistic processing features. It supports English, French and Arabic languages. LingOnto can be used by linguistically under-skilled users. The current version of LingOnto includes 216 classes, 136 object properties and 326 **S**emantic **W**eb **R**ule **L**anguage (SWRL) rules. LingOnto is never frozen, which means that we can add other linguistic knowledge.

3.1 Overview

We are based on the design principles presented by Gruber (1995), which are objective criteria for guiding and evaluating ontology designs, such as clarity, coherence, minimal encoding bias and minimal ontological commitments. Following these principles, we define the following top-level concepts of our ontology as shown in Figure 1.

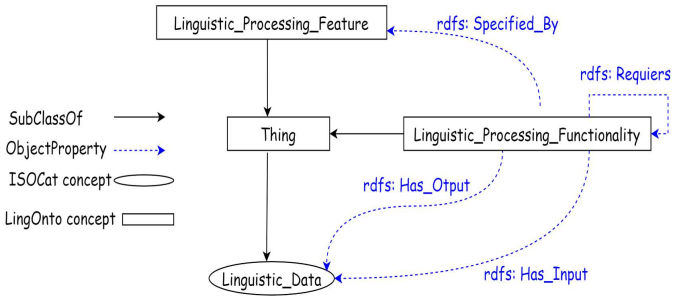

Figure 1: The top level concepts of LingOnto.

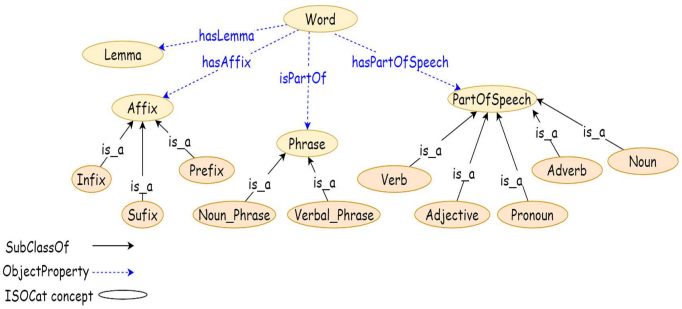

Figure 2: The Classification of some linguistic data.

3.2 Linguistic Data Classification

Referring to the ISOcat standard, we identify a set of linguistic data concepts. We choose ISOcat as it covers more terms of linguistic resources and linguistic levels compared to WordNet and GOLD. Figure 2 shows a part of LingOnto, illustrating the classification of some linguistic data. In fact, a word "Word" has a part of the speech "PartOfSpeech" which can be a noun "Noun", verb "Verb", adjective "Adjective", and so on. For this, we have defined an "is_a" object property between these latter and the "PartOfSpeech" class. Similarly, a word "Word" has an affix "Affix" which can be a prefix "Prefix", infix "Infix" or suffix "Suffix". In addition, a word "Word" is a part of a sentence "Phrase". As a consequence, an "isPartOf " object property has been established between the "Word" class (domain) and the "Phrase" class (range).

3.3 Linguistic Processing Functionality Classification

In order to identify a set of linguistic processing functionalities, a manageable selection of language

processing platforms (e.g., Grid (Ishida, 2011) and Weblicht (Hinrichs et al., 2010)) and NLP toolkits (e.g., Apache OpenNLP [3], Standford CoreNlP [4], FreeLing [5] and LingPipe[6]) is examined. The list is restricted to toolkits supporting English, French and Arabic languages. Moreover, we focus on linguistic processing functionalities, leaving out other functionalities provided by some of the toolkits. For example, FreeLing provides a variety of processors, including modules for performing tasks of statistical machine learning. In the following, we present some of the standard linguistic processors that we extract.

Linguistic Processors = {*Language Identifier, Sentence Splitter, Tokenizer, POS Tagger, Lemmatizer, Sense Tagger, Morphological Analyzer, Chunker, NE Recognizer, Coreference Resolver, Dependency Parser, Phrase Structure Parser, Speech Recognizer, TextTo-Speech Converter, Translator, Paraphraser*}

A linguistic processor implements often one or two linguistic processing functionalities. As a first example, "Morphological Analyzer" implements "Tokenization", "POS Tagging" and "Lemmatization" functionalities. As a second example, the "NE Recognizer" implements "Chunking" and "NE Classification" functionalities.

In the herein work, we intend to construct an ontology involving both lower and higher level processing functionalities in order to satisfy variable granularity user's need. Hence, we propose a set of linguistic processing functionalities as following:

Linguistic Processing Functionalities = {*Language Identification, Sentence Splitting, Tokenization, POS Tagging, Lemmatization, Sense Tagging, Morphological Analyzing, NE Recognizing, Chunking, NE Classification, Coreference Resolution, Dependency Parsing, Phrase Structure Parsing, Speech Recognition Text-To-Speech Conversion, Translation, Paraphrasing*}

After identifying a set of linguistic processing functionalities, we identify the relationships that may exist between them. There are a hierarchical interdependencies between the different linguistic processing functionalities (Hayashi and Narawa , 2012).

[3]http://opennlp.apache.org
[4]http://nlp.stanford.edu/software/ corenlp.shtml
[5]http://nlp.lsi.upc.edu/freeling/
[6]http://alias-i.com/lingpipe/

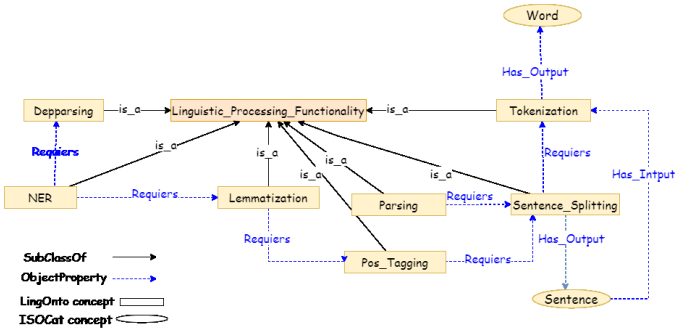

Figure 3: The Classification of some linguistic processing functionalities.

Indeed, a linguistic processing functionality used to perform analysis at one level may require, as input, the results of an analysis of a lower level. For example, syntactic analysis like parsing usually requires words to be clearly delineated and part-of-speech tagging or morphological analysis to be performed first. In the annotation process for example, the texts must be tokenized, their sentences clearly separated from each other and their morphological properties analyzed before starting the parsing functionality . Hence, we identify the object property "Requires". As shown in Figure 3, the "Tokenization" class is in relation with the "Sentence_Splitting" class through this object property. Moreover, each linguistic processing functionality manipulates various linguistic data as inputs and others as outputs. Hence, we propose the objects properties "Has_Input" and "Has_Output". For instance, as shown in Figure 3, the "Tokenization" class is in relation with the "Sentence" class through "Has_Input" object property. It is also in relation with the "Word" class through "Has_Output" object property. To reason about linguistic processing functionalities, we propose a set of SWRL rules. For example, the SWRL rules allow deducing the object property "Requires":

Linguistic_Processing_Functionality (A) $\wedge$ *Linguistic_Processing_Functionality (B)* $\wedge$ *Linguistic_data (I)* $\wedge$ *Has_Output (A,I)* $\wedge$ *Has_Input (B,I)* $\rightarrow$ *Requires (A, B).*

Linguistic_Processing_Functionality (A) $\wedge$ *Linguistic_Processing_Functionality (B)* $\wedge$ *Linguistic_Processing_Functionality (C)* $\wedge$ *Requires (B,A)* $\wedge$ *Requires (C,B)* $\rightarrow$ *Requires (C,A).*

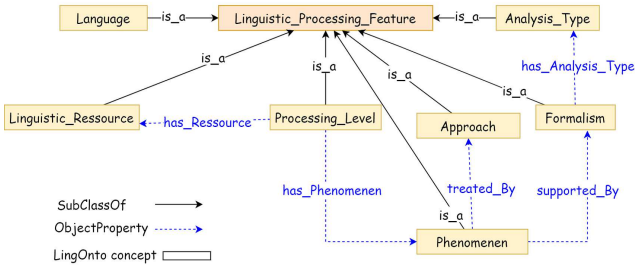

Figure 4: The Classification of some linguistic processing features.

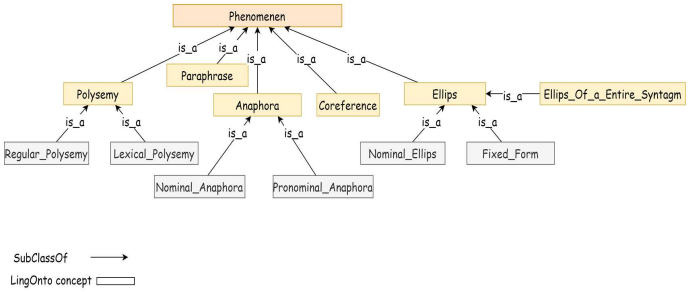

Figure 5: The Classification of some linguistic phenomenon.

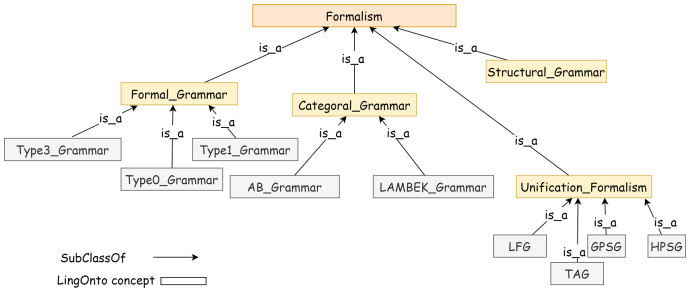

Figure 6: The Classification of some linguistic formalism.

3.4 Linguistic Processing Feature Classification

Each linguistic processing functionality is characterized by a set of linguistic processing features. We propose that the "Linguistic_Processing_Feature" class includes the following sub-classes:

- "Processing level": it represents the processing level of a linguistic processing functionality. In NLP, we distinguish mainly four processing levels: lexical, morphological, syntactic, and semantic. Each processing level is characterized by both its resources (e.g., dictionaries, tree bank and corpus) and phenomena (e.g., ellipsis, anaphora and accord). For that, we propose the object properties "has_Resource" and "has_Phenomenon".

- "Phenomenon": it is the linguistic phenomenon treated by a processing level. It has the "refined_into" object property, since each phenomenon has its subPhenomena. For example, in the ellipsis phenomenon we distinguish the nominal ellipsis (the omission of the essential part of a nominal phrase: the head) and an ellipsis of a whole phrase (e.g., subject ellipsis, verb ellipsis, both verb and complement ellipsis). The "Phenomenon" class also has a "treated_By" object property in relation with the "Approach" class and "supported_By" object property with the "Formalism" class.

- "Approach": it is the linguistic approach treated by a phenomenon. It can be a statistical, linguistic or hybrid approach (linguistic and statistical).

- "Formalism": it is the linguistic formalism that supports a phenomenon. There are several types of formalisms such as HPSG and LFG for syntactic grammars (Kahane, 2006). A formalism can concern a main phenomenon or a subphrenomen. Also, it can have a type of analysis. So, we propose the "has_Analysis_Type" object property with the "Analysis type" class.

- "Analysis type": it is the type of analysis that characterizes a linguistic formalism, namely, bottom-up analysis, top-down analysis, surface analysis and so on.

- "Language": it is important to learn about the specificity and the structure of each language to deal with its complexity. For example, Arabic language is a very rich language with complex morphology, with different and difficult structure than other languages. LingOnto focus on English, French and Arabic languages.

4 Experimentation

We apply the proposed ontology to a framework of identifying valid composition workflows of LingWS

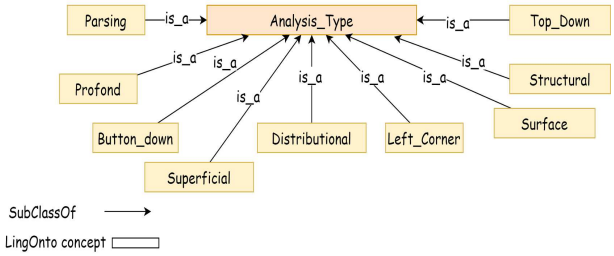

Figure 7: The Classification of some linguistic analysis type.

(Neji et al., 2018). Then, we evaluate its efficiency in the context of the last one.

4.1 Application to LingWS Composition Workflows Identification Framework

The LingOnto is applied to a framework of identifying valid composition workflows of LingWS. It targets under-skilled users in the lingware engineering area.

First, the user generates a dynamic ontological view from LingOnto based on a set of selection criteria. This is done thanks to a user friendly ontology visualization tool called LingGraph (Neji et al., 2019). The choice of one or more criteria is made to show only the components corresponding to the user's need. For instance, Figure 8 shows LingOnto's components representing the different linguistic processing functionalities related to the morphological level of English language. Second, the user starts the identification of a workflow of linguistic processing functionalities based on the generated ontological view. If the user selects a functionality, LingOnto proposes a set of possible functionalities choices that can be added to the workflow. This step is done thanks to the aforementioned LingOnto SWRL rules. For instance, as shown in Figure 8, a Part-of-Speech Tagging functionality can be added to the workflow only after a Tokenization functionality. Finally, the corresponding LingWS(s) to each selected linguistic processing functionality is discovered taking into account the set of linguistic processing features presented in LingOnto. Indeed, the discovery process performs a matching between the linguistic processing features of each selected linguistic processing functionality and the description of each required LingWS. This step explores the LingWS registry (Baklouti et al., 2015).

4.2 Evaluation

A total of 30 users were recruited to participate in this evaluation study. They are researcher members of NLP Research Group of MIRACL laboratory (Tunisia, Sfax) and CEDRIC laboratory (France, Paris). The selected users have the same NLP and languages competences. Before beginning the experiment, they were asked to fill a pre-questionnaire about their prior knowledge and expertise in NLP research field and languages. These users are equally allocated into three groups where each group focus on only one language i.e., English, French or Arabic. Each user identified a set of composition workflows from LingOnto related to the morphological level. An NLP domain expert participated in this evaluation in order to identify the number of valid possible composition workflows corresponding to the morphological level of each language.

For each group of users working on a given language, we note:

- All_User_W : the total number of composition workflows identified by all the users of the group without redundancy.
- V_All_User_W : the total number of **valid** composition workflows identified by all the users of the group without redundancy.
- User_W : the total number of composition workflows identified by a given user of the group.
- V_User_W : the total number of **valid** composition workflows identified by a given user of the group.
- Exp_W : the total number of **valid** composition workflows identified by the NLP domain expert for the concerned language.

We use the Recall and Precision evaluation metric as follow:

- The Recall associated to a given language R_L = *(V_All_User_W / Exp_W)*.
- The Precision associated to a given language P_L = *(V_All_User_W / All_User_W)*.

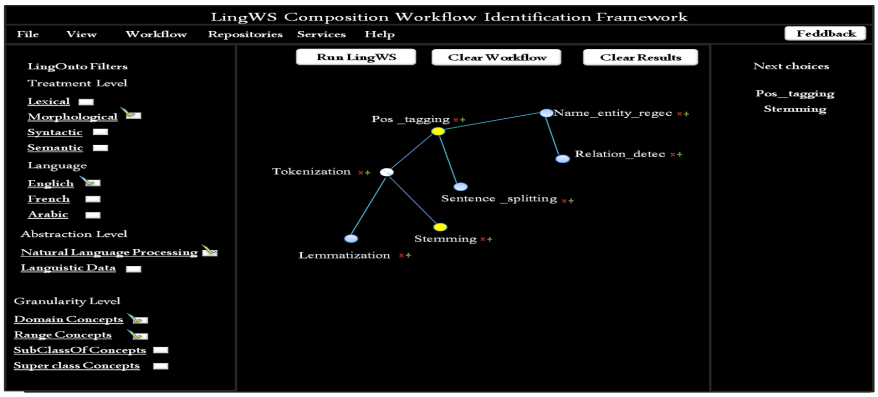

Figure 8: Screenshot of LingWS composition workflows identification framework showing a part of LingOnto.

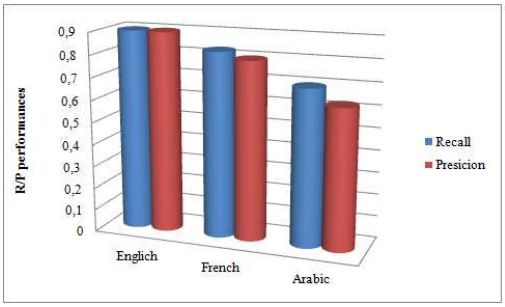

Figure 9: R/P performances of the three languages.

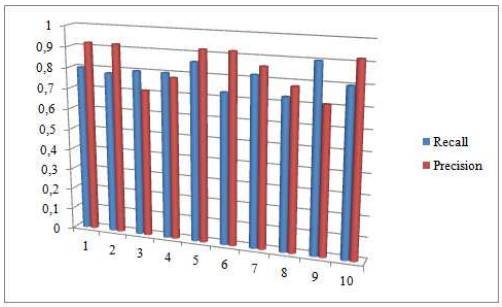

Figure 10: R/P performances of English users.

- The Recall associated to a given user R_U = *(V_User_W / Exp_W)*.
- The Precision associated to a given user P_U = *(V_User_W / User_W)*.

As shown in Figure 9, the mean of the R/P performances measures indicates that our ontology is efficient in identifying valid composition workflows. Besides, the overall mean of the precision associated to the English and French language is better than the overall mean of the precision associated to the Arabic language. This gap of R/P performances is explained by the fact that the Arabic language is characterised by a complicated morphology compared to English and French languages. Figure 10 shows the R/P performances of the English language users. We note that these performances are not the same for all the users. While the users engaged for the English language have similar levels of competences, then the difference in their performance could reflect a complexity in exploiting the visualization tool LingGraph.

5 Conclusion and Future Work

In this paper, we proposed a multilingual linguistic domain ontology, called LingOnto, for representing and reasoning about linguistic knowledge. It helps linguistically under-skilled users in understanding the meaning, scope and usage of linguistic knowledge.

At the beginning, we elaborated a state of the art focusing on approaches that are proposed to represent linguistic knowledge. This study showed that existing works consider only linguistic data.To the best of our knowledge, there is no approach that allows representing and reasoning about linguistic processing functionalities and features and their linking with linguistic data. Moreover, most of them do not support multilingualism. Compared to related work, LingOnto allows representing and reasoning about linguistic data and linguistic processing functionalities and features. It supports English, French and Arabic languages. We applied LingOnto to a framework of identifying valid composition work-

flows of LingWS.

Currently, we are applying LingOnto to a smart memory prosthesis for Alzheimers patients called CAPTAIN MEMO. It is proposed in the context of VIVA project *(Vivre a Paris avec Alzheimer en 2030 grâce aux nouvelles technologies)*. LingOnto is used to identify composition workflows of a nature language-interrogation system that aims to facilitate the communication between the prosthesis and the Alzheimer's patients.

We plan to allow the LingOnto ontology to be referenced by the Linked Open Vocabularies (LOV) platform. Moreover, we plan to exploit the NLP domain expert's feedback to improve the proposed ontology.

References

Baklouti, N., Bouaziz, S., Gargouri, B., Aloulou, C., and Jmael, M. 2010. *Towards the reuse of lingware systems: a proposed approach with a practical experiment*. International Conference on Information Integration and Web-based Applications and Services, pages 566572.

Baklouti, N., Gargouri, B., and Jmaiel, M. 2015. *Semantic-based approach to improve the description and the discovery of linguistic web services* . Engineering Applications of Artificial Intelligence, 46, 154-165.

Chiarcos, C. and Sukhreva, M. 2015 . *OLIA - Ontologies of Linguistic Annotation*. Semantic Web Journal, 518:379386.

De Cea, G. A., de Mon, I. A., Gomez-Perez, A., and Pareja-Lora, A. 2004. *Ontotags linguistic ontologies: improving semantic web annotations for a better language understanding in machines*. International Conference on Information Technology: Coding and Computing(ITCC). (Vol. 2, pp. 124-128). IEEE.

Eugene E. Loos, Susan Anderson, Dwight H. Day, Jr., Jordan Paul. 2004. *Glossary of linguistic terms, volume 29*. SIL International.

Fellbaum, C. 2012. *WordNet*. The encyclopedia of applied linguistics.

Farrar, S., and Langendoen, D. T. 2010. *An owl-dl implementation of gold: An ontology for the semantic web* . Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology.

Gangemi A., Guarino, N., Oltramari, A. and Borgo, S. 2002. *Cleaning-up wordnets top-level*. Proceedings of the 1st International WordNet Conference.

Gruber, T. R. 1995. *Toward principles for the design of ontologies used for knowledge sharing?* . International journal of human-computer studies, 43(5-6), 907-928.

Hayashi, Y., and Narawa, C. 2012 . *Classifying Standard Linguistic Processing Functionalities based on Fundamental Data Operation Types*. LREC (pp. 1169-1173).

Hinrichs, M., Zastrow, T., and Hinrichs, E. W. 2010. *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure*. International Conference on Language Resources and Evaluation, pages 489493,Valletta, Malta.

Ishida, T. (Ed.). 2011. *The language grid: Service-oriented collective intelligence for language resource interoperability*. Springer Science and Business Media.

Kahane, S. . *Polarized unification grammars* . International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 137-144.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. E. 2009. *Isocat: Remodeling metadata for language resources*. International Journal of Metadata, Semantics and Ontologies, 4(4), 261-276.

Neji, M., Gargouri, B., and Jmaiel, M.. 2018. *A semantic approach for constructing valid composition scenarios of linguistic Web services*. Procedia Computer Science, 126, 685-694.

Neji, M., Ghorbel, F., and Gargouri, B. 2019. *A smart search-based ontology visualization tool using sparql patterns*. International Conference on Knowledge Science, Engineering and Management (pp. 33-44). Springer, Cham.

Schuurman, I., Windhouwer, M., Ohren, O., and Zeman, D. 2016. *Clarin concept registry: the new semantic registry*. Selected Papers from the CLARIN Annual Conference 2015, pages 6270.

Zhou, M., Duan, N., Liu, S., and Shum, H. Y.. 2020. *Progress in neural nlp: Modeling, learning, and reasoning*. Engineering, 6(3), 275-290.