HOLINET: Holistic Knowledge Graph for French

Jean-Philippe Prost

HOLINET: Holistic Knowledge Graph for French
Graphe de connaissances holistique pour le français

Jean-Philippe Prost[1]
(1) Laboratoire Parole et Language, CNRS – Aix-Marseille Université, France
Jean-Philippe.Prost@univ-amu.fr

RÉSUMÉ

HOLINET est un graphe de connaissances pour le français, qui vise à fournir une perspective holistique de la représentation des connaissances linguistiques. Ainsi, il approche la langue à la fois comme tout, et comme la somme de ses parties sur les différentes dimensions linguistiques. Nous formulons l'hypothèse qu'une telle modélisation holistique des connaissances linguistiques facilitera le traitement automatique du language et améliorera les performances des applications en aval. HOLINET ouvre de nouvelles pistes de recherche en tant que graphe de connaissances qui intègre des connaissances syntaxiques de référence et des connaissances lexico-sémantiques, et qui combine constituance et dépendance. L'encodage de HOLINET en un modèle de plongement de graphe de connaissances reste une perspective saillante à explorer.

ABSTRACT

HOLINET : Holistic Knowledge Graph for French

HOLINET is a knowledge graph (KG) for French, which aims to provide a holistic perspective on language knowledge representation. As such, it approaches language as a whole, as well as a sum of its parts on various linguistic dimensions. We hypothesize that a holistic modelling of language knowledge will ease its processing and improve the performance of downstream applications. HOLINET opens new avenues of research as a KG which integrates gold-standard syntactic knowledge along with lexical semantic one, and which is open to combining constituency and dependency information. The computation of a KG embedding model, for instance, is a salient option to investigate.

MOTS-CLÉS : Graphe de connaissances, réseau lexico-sémantique, grammaire syntagmatique.

KEYWORDS: Knowledge Graph, Lexical-Semantic Network, Phrase Structure Grammar.

1 Introduction

Knowledge Graphs (KGs) have become a corner stone of modern Artificial Intelligence (AI), for the central role they play in a variety of downstream applications, such as QA, recommender systems, semantic parsing, etc. They suit both symbolic and sub-symbolic types of processing, since they may be involved in these applications in plain form, as part of graph-theoretical processes (e.g. path-based reasoning), or in sub-symbolic form, where the graph is converted to a numerical representation, such as knowledge graph embeddings.

As far as NLP applications are concerned, KGs are often relied on for lexical and semantic knowledge, usually through KG emmbeddings, while syntax is usually gathered from other sources, typically

annotated corpora. In this case, the integration between syntax and lexical semantics is done by the application. But what if the integration was done earlier, in the KG ? Would embeddings encoded on such a KG perform better than the existing models for the relevant NLP tasks (e.g. semantic parsing) ?

More generally, while the pipeline software architecture, which steps from one linguistic dimension to the next, has been typical for decades for most NLP applications, it often prevents many potential interactions across dimensions from actually occurring. A variety of sentence-level ambiguities, for instance, require the full sentence to be parsed morphologically, then syntactically, then semantically, prior to being disambiguated through a pipeline. Knowledge graphs provide a convenient means for heterogeneous knowledge to interact rather seamlessly.

Prior to addressing questions such as the performance of syntax-semantic Knowledge Graph embeddings in NLP tasks, this work focuses on the construction of such an integrated KG, and the problems that come along with it. Section 2 introduces some background knowledge and review the literature. Section 3 presents the graph model underpinning the HOLINET knowledge graph and its implementation, along with evaluation figures. Section 4 discusses further works, and section 5 concludes.

2 Background and literature review

What is a Knowledge Graph ? As Hogan *et al.* (2021, p. 2) put it, "[t]he definition of a "*knowledge graph*" remains contentious". Their own definition is an inclusive one,

> "(. . .) where we view a knowledge graph as a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities. The graph of data (aka data graph) conforms to a graph-based data model, which may be a directed edge-labelled graph, a property graph, etc. (. . .). By knowledge, we refer to something that is known".

Literature review Knowledge Bases for natural language, whether structured as graphs or not, are rarely holistic, in the sense that they would merge, and present, all linguistic dimensions as a whole, within a single and homogeneous structure. The Linguistic Linked Open Data (LLOD) initiative links together different resources. LLOD's interest goes towards representation format problems, or federation of multiple data sources, or interoperability. But the seamless integration of heterogeneous data is not its prime concern – especially not so at the interface between syntax and lexical semantics. In fact, as far as we know, LLOD only links syntactic resources through annotated corpora. We do not know of any grammatical knowledge base, regardless of the language. Faralli *et al.* (2020) is concerned with integrating 5 resources already linked through LLOD : ConceptNet (Speer *et al.*, 2017), DBpedia (Lehmann *et al.*, 2015), WebIsAGraph (Faralli *et al.*, 2019), WordNet (Miller, 1995) and the Wikipedia category hierarchy. No syntactic resource involved.

The lexico-semantic graphs relie on various structures. WordNet is a network of 150K+ words, organised in 170 000 synsets, which can be seen as concepts. WOLF (Sagot & Fier, 2008) is a French version of it. The French Lexical Network (*Réseau Lexical du Français*) (RLF) from Polguère (2014) relies on the notion of *lexical function* as defined by Igor Mel'čuk. JeuxDeMots (JDM) is another lexico-semantic network for French (Lafourcade, 2007), which implements the same notion and generalises it beyond lexical knowledge. JDM is made up of 16,5+ millions nodes, including

5,2+ millions terms, 400+ millions relationships and 150+ relationship types. The words are terms, concepts and symbolic information. The relationship types are lexical, morphological, pragmatical, logical, ontological, etc.

None of these resources include grammatical knowledge. FrameNet (Baker *et al.*, 1998), which is dedicated to *frame semantics* (Fillmore, 2008), relates the semantic frames with each others through semantic relationships to constitute a network. Each frame is illustrated with prototypical utterances, annotated with syntax. However, the syntactic analysis is expressed through text annotations, and is not, strictly speaking, integrated in the network.

The literature shows that syntactic knowledge is mainly represented through annotated resources, and to some extent through symbolic grammars, such as the HPSH grammars (Pollard & Sag, 1994) from the DELPH-IN consortium (Copestake & Flickinger, 2000), or meta-grammars such as FRMG (de La Clergerie, 2005).

3 What is HOLINET about ?

HOLINET aims to provide a holistic perspective on language knowledge representation. Holistic in that all linguistic dimensions, although heterogeneous by nature, are integrated within the same data structure. As such, it approaches language as a whole, as well as a sum of its parts on various dimensions. One of the main motivation for such an approach is to overcome some of the issues raised by the traditional pipeline architecture for NLP applications.

The HOLINET graph model has already been thoroughly detailed in (Prost, 2022), along with its automated construction process [1] [2]. The resource is original in that it integrates lexical semantic knowledge and grammar knowledge within a single graph. The lexical semantic layer is the lexico-semantic network JeuxDeMots (JDM) (Lafourcade, 2007). It conveniently represents the POS categories of all the terms in the network. conveniently, because the POS categories will serve as the interface between JDM and the grammar layer to come. The grammar layer is made up of a phrase structure grammar, which may be seen as a set of context-free rewrite rules, such as Rule (1) illustrated in Figure 1. The POS in a rule which pre-exist in JDM are as many anchor nodes. The phrasal categories, e.g. Noun Phrase (NP) or Adjective Phrase (AP), do not pre-exist in JDM, hence are created for the grammar layer in HOLINET. Note that the type `n_pos` in HOLINET generalises both the actual POS categories and the phrasal categories.

Figure 1 illustrates the graph model of the grammar layer for the example rule (1). The POS categories are pre-existing nodes of the JDM type `n_pos`. The terms are related to their respective POS categories with pre-existing relationships of the JDM type `r_pos`. Every rewrite rule is reified as a node, typed `n_g_cfRule`. The left-hand side of each rule is itself reified as a POS node (of type `n_pos`), and every rule is connected to its left-hand side POS node with the `r_g_rewrites` relation. Meanwhile, on the right-hand side (RHS) of each rule, every constituent POS is connected to its rule with the `r_g_constitutes` relation. In order to allow redundancy of POS, like here the AP, every constituent on the RHS is related to its POS node with an `r_g_instantiates` relation. The feature structure next to Rule (1) details the properties associated with the `n_g_cfRule` node

1. HOLINET v1.0 is distributed by ORTOLANG (`https://hdl.handle.net/11403/holinet-1-0/v1`) under a Creative Commons Attribution 4.0 International licence (CC-BY 4.0).

2. All the software involved in the creation process is publicly available as git repositories on sourceforge.net. See (Prost, 2022) for more details.

(1) NP →DET#fsi NC#fsc AP AP

$$\text{RULE (1)}\left[\begin{array}{l}\text{TYPE}\ \texttt{n\_g\_cfRule}\\ \text{RULE}\left[\begin{array}{l}\text{LHS}\ [\text{CAT}\ NP]\\ \text{RHS}\left\langle \underset{\text{DET\#FSI}}{\left[\begin{array}{ll}\text{CAT} & DET\\ \text{CPOS} & D\\ \text{GEND} & f\\ \text{NUM} & s\\ \text{SUBCAT} & ind\end{array}\right]}, \underset{\text{N\#FSC}}{\left[\begin{array}{ll}\text{CAT} & NC\\ \text{CPOS} & N\\ \text{GEND} & f\\ \text{NUM} & s\\ \text{SUBCAT} & c\end{array}\right]}, [\text{CAT}\ AP], [\text{CAT}\ AP]\right\rangle\end{array}\right]\end{array}\right]$$

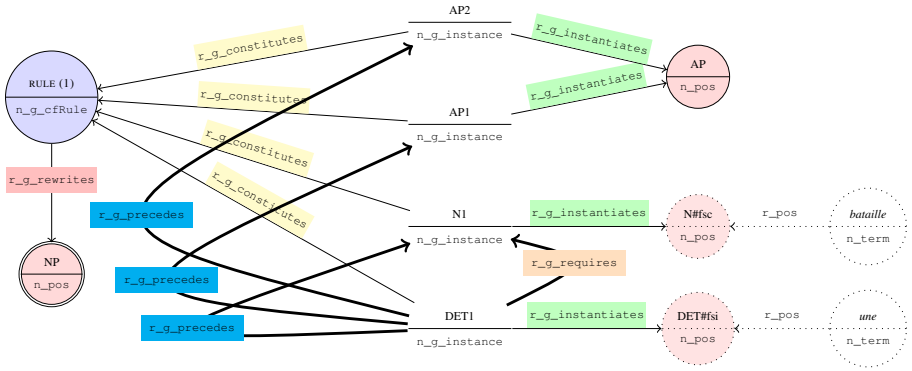

FIGURE 1 – HOLINET Graph model for the example context-free grammar rewrite rule (1). *The dotted nodes and edges originate in JDM. They serve as anchorage for connecting the grammar layer to JDM.*

labelled RULE (1).

Next to the Immediate Dominance relationship, represented by the `r_g_constitutes` relations, the model also represents other relationships (in bold), such as Linear Precedence (`r_g_precedes`), and Requirement (`r_g_requires`). Their semantics is borrowed to the corresponding well-defined *properties* in Property Grammar (PG) (Blache, 2001), a formal framework for specifying constraint-based grammars. Further works will investigate the integration of even more relationships, such as *obligation* (to model a phrase's head), *agreement* and *dependency* (for dependency grammars).

The creation process in short For the sake of generalisation, we assume (a) a constituency treebank, and (b) a lexical network. The lexical entries in the network are expected to be related to their POS categories, where the relationship is of type `r_pos`, and the POS are as many nodes. If necessary, we assume a mapping between the two POS tagsets, i.e. the network's and the treebank's. The creation process is, then, made up of the following steps :

1. *read/extract* the implicit CFG from the constituency trees in the treebank
2. *derive* the additional relationships from the CFG (linear precedence, requirement, etc.)
3. *assess* the truth values to be assigned to the relationships (default is all true).
   Given a corpus CFG, the process of derivation (step 2), then assessment (step 3) of the relationships we are interested in, is equivalent to the so-called "compilation process" of a Property Grammar, as described by Prost *et al.* (2016).
4. *map* the treebank tagset to the network's (or the other way around)

5. *create* the required grammar triples (i.e., nodes and relations), according to the augmented CFG (i.e., with the additional relationships from step 2)

6. *merge* the two layers and check their consistency (detailing this step further goes beyond the scope of this paper).

Experiments and first evaluation The creation process has been implemented with the version 1.0 (2016) of the FTB annotated in the Penn Treebank format, and the version of JDM extracted from the dump dated 01/11/2021. But we believe that the process is trivial enough to be easily adapted to other resources.

Since we are primarily concerned with integrating the grammar layer with the lexical-semantic layer, and since the anchorage of the grammar layer to JDM is achieved through the POS nodes shared by the two layers, we want to measure the proportion of the POS categories required by the grammar layer (i.e., found in the FTB) that are actually found in JDM. We are only interested in the actual POS, not the phrase categories, in spite of the fact that both are modelled in HOLINET as nodes of type `n_pos`. Our evaluation is reported in Table 1. It shows that only 30.1% of the POS categories

	Connected	Disconnected	Null	**Total**
Num. nodes	22,742	43361	9,408	**75511**
%	30.1%	57.4%	12.5%	**100**

TABLE 1 – proportion of the POS categories found in FTB grammar rules, that are actually found in JDM (connected), or not (disconnected). The phrase categories, although typed `n_pos`, have been taken out of the picture. The 'Null' value stands for mapping issues.

involved in the grammar are actually found *as such* in JDM.

Most of the discrepancies between the two layers come from the choice JDM makes to split among several nodes the different attributes associated with a category (e.g. gender, number, etc.), while on the grammar layer a single node is required. For instance, the FTB tag `P+PRO##cpos=P+PRO|g=f|n=p|p=3|s=rel##` is, theoretically, mapped to the JDM label `Pre+Pro:Fem+PL+Rel`. The expected corresponding labelled node is actually absent from JDM, but the following nodes typed `n_pos` are present : `Pre:`, `Pro:Fem+PL`, `Pro:Rel`, and `Pre:`. That is, all the required information is present, but split across distinct nodes.

Note that coming up with an exact algorithm, which would create the required merged nodes from the distinct ones is not trivial. Take, for instance, the term 'les'. It is connected in JDM to the following distinct `n_pos` nodes : `Det:`, `Pro:`, `Pro:Pers:COD`, `Pro:Pers`, `Pro:PL+P3`, `Det:Fem+PL`, `Det:Mas+PL`, `Det:InvGen+PL`, `Gender:Mas`, `Number:Plur`, `Gender:Fem`, `Defini:`. Quite obviously computing all the combinations is an option that would not make much sense. Improving the mapping between JDM and the HOLINET grammar layer is, therefore, ground for further investigation.

4 Applications and further works

Would a KG embedding model computed from HOLINET with both semantic and syntactic relationships improve downstream applications, such as semantic parsing ? The injection of syntactic knowledge in neural models for semantic parsing, whether deep or shallow, has been consistenly shown to improve performance. Roth & Lapata (2016) use a dependency path embedding model to improve a Recurrent Neural Network model for Semantic Role Labelling (SRL). Wang *et al.* (2019) show that the injection of syntax as input features into three different neural SRL encoders significantly improves performance. Their works also show that constituency features perform best, ahead of dependency and categorical constituency spans. Xu *et al.* (2018) combine word order, dependency and constituency features within graph embeddings. More recent works by Fei *et al.* (2021) succesfully investigate the combination of constituency and dependency through TreeLSTM and Graph Convolutional Network. Kurtz *et al.* (2019) suggest that only gold-standard syntactic information, as opposed to automatically predicted one, improves the performance of a deep neural architecture for semantic parsing. Moreover, the integration of syntactic knowledge along with lexical and semantic knowledge within the same embeddings is modern ground for investigation (Limisiewicz & Mareček, 2020; Al-Ghezi & Kurimo, 2020). In line with this body of work, HOLINET opens new avenues of research as a KG which integrates gold-standard syntactic knowledge along with lexical semantic one, and which is open to combining constituency and dependency information. The computation of a KG embedding model is, then, a salient option to investigate.

More avenues of research

— Could HOLINET integrate other types of grammar knowledge, such as dependency grammar, or Construction Grammar, and how ?

— Could the interaction between syntactic and semantic knowledge be captured in deductive and/or inductive reasoning processes for link prediction ? For desambiguation ?

5 Conclusion

In this paper we investigated the question of the integration of grammar knowledge and lexical semantic knowledge within a homogeneous graph structure, in order to construct a holistic knowledge graph for French. Our motivation is to implement an environment that enables the investigation of integrated syntax-semantic knowledge graph embeddings and their performance in downstream applications, or graph-theoretical algorithms for automated reasoning.

We presented a graph model for a phrase structure grammar, and we showed how to merge it with a lexical semantic network through a shared tagset for POS categories. We experimented the creation procedure with the French treebank (FTB) annotated for constituency, and the lexico-semantic network JeuxDeMots (JDM). Our evaluation shows that 30.1% of the POS required by the FTB can actually be found in JDM as a single node. This figure does not jeopardize the graph model as such, but rather shows that, although all the required information can be found in JDM, further work is still necessary in order to better map the annotation schemes.

Références

AL-GHEZI R. & KURIMO M. (2020). Graph-based syntactic word embeddings. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, p. 72–78.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, p. 86–90.

BLACHE P. (2001). *Les Grammaires de Propriétés : des contraintes pour le traitement automatique des langues naturelles*. Hermès Sciences.

COPESTAKE A. & FLICKINGER D. (2000). An open source grammar development environment and broad-coverage English grammar using HPSG.

DE LA CLERGERIE É. (2005). From Metagrammars to Factorized TAG/TIG Parsers. In *Proceedings of IWPT'05 (poster)*, Vancouver, Canada.

FARALLI S., FINOCCHI I., PONZETTO S. P. & VELARDI P. (2019). Webisagraph : A very large hypernymy graph from a web corpus. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, p. 13–15, Bari, Italy.

FARALLI S., VELARDI P. & YUSIFLI F. (2020). Multiple knowledge graphdb (mkgdb). In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 2325–2331.

FEI H., WU S., REN Y., LI F. & JI D. (2021). Better combine them together ! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 549–559 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.49.

FILLMORE C. J. (2008). *Cognitive Linguistics : Basic Readings*, chapitre Frame semantics, p. 373–400. De Gruyter Mouton. DOI : doi :10.1515/9783110199901.373.

HOGAN A., BLOMQVIST E., COCHEZ M., D'AMATO C., MELO G. D., GUTIERREZ C., KIRRANE S., GAYO J. E. L., NAVIGLI R., NEUMAIER S. *et al.* (2021). Knowledge graphs. *ACM Computing Surveys (CSUR)*, **54**(4), 1–37.

KURTZ R., ROXBO D. & KUHLMANN M. (2019). Improving semantic dependency parsing with syntactic features. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, p. 12–21, Turku, Finland : Linköping University Electronic Press.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition. In *Proc. SNLP 2007*, p. 13–15, Pattaya Thailande, December. 8 p : 7th Symposium on Natural Language Processing.

LEHMANN J., ISELE R., JAKOB M., JENTZSCH A., KONTOKOSTAS D., MENDES P. N., HELLMANN S., MORSEY M., VAN KLEEF P., AUER S. *et al.* (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, **6**(2), 167–195. DOI : 10.3233/SW-140134.

LIMISIEWICZ T. & MAREČEK D. (2020). Syntax Representation in Word Embeddings and Neural Networks–A Survey. arXiv preprint arXiv :2010.01063.

MILLER G. A. (1995). WordNet : a lexical database for English. *Communications of the ACM*, **38**(11), 39–41.

POLGUÈRE A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, **27**(4), 396–418.

POLLARD C. & SAG I. (1994). *Head-driven Phrase Structure Grammar*. University of Chicago Press.

PROST J.-P. (2022). Integrating a phrase structure corpus grammar and a lexical-semantic network : the holinet knowledge graph. In *Proceedings of LREC 2022 the 13th Language Resources and Evaluation Conference*, p. 613–622, Marseille, France : European Language Resources Association European Language Resources Association. HAL : hal-03655636.

PROST J.-P., COLETTA R. & LECOUTRE C. (2016). Compilation de grammaire de propriétés pour l'analyse syntaxique par optimisation de contraintes. In *Actes de TALN 2016, 23ème conférence sur le Traitement Automatique des Langues Naturelles*, p. 396–402, Paris, France : Association pour le Traitement Automatique des Langues.

ROTH M. & LAPATA M. (2016). Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1192–1202 : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1113.

SAGOT B. & FIER D. (2008). Construction d'un WordNet libre du français à partir de ressources multilingues. In *Proceedings of TALN 2008*, Avignon, France.

SPEER R., CHIN J. & HAVASI C. (2017). Conceptnet 5.5 : An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

WANG Y., JOHNSON M., WAN S., SUN Y. & WANG W. (2019). How to best use syntax in semantic role labelling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5338–5343 : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1529.

XU K., WU L., WANG Z., YU M., CHEN L. & SHEININ V. (2018). Exploiting rich syntactic information for semantic parsing with graph-to-sequence model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 918–924 : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1110.