Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5163–5169
Marseille, 20-25 June 2022

Introducing RezoJDM16k: a French Knowledge Graph DataSet for Link Prediction

Mehdi Mirzapour[1], Waleed Ragheb[2,5], Mohammad Javad Saeedizade[3],
Kévin Cousot[4], Hélène Jacquenet[1], Lawrence Carbon[1], Mathieu Lafourcade[2]

ContentSide[1]; LIRMM, Univ Montpellier[2]; IUST[3]; Emvista[4]; FCAI, Cairo Univ[5]
R&D Dept, Lyon, France[1]; CNRS, Montpellier, France[2]; Tehran, Iran[3]
R&D Dept, Montpellier, France[4]; Cairo, Egypt[5]
{first.last}@contentside.com[1]; {first.last}@lirmm.com[2]
m_saeedizade@comp.iust.ac.ir[3]; kevin.cousot@emvista.com[4]

Abstract

Knowledge graphs applications, in industry and academia, motivate substantial research directions towards large-scale information extraction from various types of resources. Nowadays, most of the available knowledge graphs are either in English or multilingual. In this paper, we introduce RezoJDM16k, a French knowledge graph dataset based on RezoJDM (Lafourcade, 2007). With 16k nodes, 832k triplets and 53 relation types, RezoJDM16k can be employed in many NLP downstream tasks for the French language such as machine translation, question-answering and recommendation systems. In addition, we provide strong knowledge graph embedding baselines that are used in link prediction task for future benchmarking. Compared to the state-of-the-art English knowledge graph datasets used in link prediction, RezoJDM16k shows a similar promising predictive behavior.

Keywords: language resource, knowledge graph dataset, link prediction, knowledge graph embedding, knowledge graph completion, lexical-semantic network

1. Introduction

Knowledge Graphs (KGs) are structured representations of semantic information mainly used for different tasks in artificial intelligence such as information extraction, search engines, question answering, and recommendation systems. KGs are often represented as multi-relational graphs with nodes and different types of edges. In a KG, each link is a triplet of the form (head, relation, tail). A triplet is a semantic representation of external world fact in which head and tail are entities. They are connected by a relation which acts as the semantic predicate between the entities. For instance the triplet (*hunt* $\xrightarrow{r\_agent}$ *lion*) indicates that *lion* is the agent of verb *hunt*, and the (*coffee* $\xrightarrow{r\_carac}$ *hot*) indicates that *coffee* has the characteristic of being *hot*. There is no limitation to have a large number of facts collected in KGs.

Although the basic idea behind KGs seems very promising, there is a crucial problem about KGs that makes them challenging to utilize: they are always incomplete (Wang et al., 2021). We can observe a lot of missing information (links or relations) between the entities in KGs. Moreover, real-world data are often dynamic and evolving, which makes it hard to build complete KGs (Cai et al., 2018; Arora, 2020). This is the reason behind the necessity of predicting missing information in KGs to make them as complete as possible. This task is called *Link Prediction* or *Graph Completion* which semantically refer to a unique notion. The most successful approach to address *Link Prediction* problem is based on *Knowledge Graph Embedding* (KGE) methods which transform KGs into a low-dimensional vector space. This transformation, in principle, should preserve the structure of the KG and their underlying semantics (Wang et al., 2021).

There are famous KGs such as Freebase (Bollacker et al., 2008), DBpedia (Lehmann et al., 2015) and WordNet (Miller, 1995) which contain huge number of entities and relations. As for Freebase, there are currently around 3.1 billion triplets and more than 110 million entities. Some studies show that working with huge KGs can impact the quality and interpretability of the evaluations (Socher et al., 2013). This suggests a kind of filtering or graph subselection algorithm. For instance, we can mention FB15k and FB1M datasets (Bordes et al., 2013) which are created by selecting the most frequently occurring of triplets in Freebase KG. The FB15k dataset suffered from major test leakage through inverse relations, where several test triplets could be obtained by inverting triplets in the training set. As a result, another subset of FB15k, which is called FB15k-237, introduced in (Toutanova et al., 2015). The same practice is done for creating WN18RR, which is a subset of dataset WN18 created from WordNet (Toutanova et al., 2015).

The above-mentioned KG datasets have been widely accepted for the English language. But, to the best of

our knowledge, there is no efficient proposal to create KG datasets for the French language. There are some multilingual KGs such as ConceptNet (Speer et al., 2017), BabelNet (Navigli and Ponzetto, 2012) which can be partially used in order to create a French KG. Nevertheless, working with specialized French large lexical-semantic network seems a more promising strategy. For doing such a task, we have focused our study on RezoJDM, which is a lexical-semantic network for the French language (Lafourcade, 2007). It contains commonsense knowledge that is lacking in ConceptNet and BabelNet (Lafourcade and Le Brun, 2017). Any KG, such as ConceptNet, that only focuses on concept can potentially miss important relations. In contrast, lexico-semantic approach combines lexical level and conceptual level information. Also, ConceptNet has no explicit representation of polysemy which is the case in RezoJDM (Chatzikyriakidis et al., 2017). As for BabelNet, the French side has some errors, mainly due to the automatic approach of linking the English and the French entities through machine translation. Building RezoJDM is performed by crowd-sourcing through several games with a purpose[1] (GWAPs), direct contributions[2] and a set of automatic inference processes.

RezoJDM aims at providing general lexical and semantic knowledge with a strong focus on common sense. The network's nodes represent any type of lexical item from single words (such as *chair*) to more complex expressions (such as *to sit on a chair*). Edges are typed so to express a particular relationship between two lexical items. Relationship types can be divided into different categories: lexical (synonymy, antonymy, ...), ontological (hyperonymy, meronymy, ...) and predicative (agent, consequence, ...). Tables (1) and (2) show the definition and typical examples of most frequent relation types. RezoJDM currently has around 5.2 millions nodes, 400 millions edges and 140 relationship types.

Type	Description
r_agent	Entity that performs the action
r_patient	Entity that undergoes the action
r_carac	Object's characteristic
r_causatif	Possible cause
r_conseq	Possible consequence
r_has_part	Whole to part
r_holo	Part to whole
r_instr	Action's instrument
r_isa	Specific to general
r_lieu	Typical place

Table 1: Descriptions of relation types in RezoJDM

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

[2]http://www.jeuxdemots.org/diko.php

Type	Example
r_agent	hunt $\xrightarrow{r\_agent}$ lion
r_patient	hunt $\xrightarrow{r\_patient}$ antelope
r_carac	coffee $\xrightarrow{r\_carac}$ hot
r_causatif	hunting $\xrightarrow{r\_causatif}$ hunger
r_conseq	hunger $\xrightarrow{r\_conseq}$ eat
r_has_part	house $\xrightarrow{r\_has\_part}$ room
r_holo	room $\xrightarrow{r\_holo}$ house
r_instr	fishing $\xrightarrow{r\_instr}$ fishing rod
r_isa	mammal $\xrightarrow{r\_isa}$ animal
r_lieu	Times Square $\xrightarrow{r\_lieu}$ New York

Table 2: Examples of relation types in RezoJDM

In this research, we mainly focused on the creation of an efficient French KG dataset that can straightforwardly be fed into current state-of-the-art KGE models. Such models for the French language can be used in different tasks such as predicting missing information, recommender systems, question answering, query expansion, etc. As discussed, we explore different sub-graph selection criteria to make an efficient algorithm to get the most informative part of RezoJDM. We also provide some predictive model baselines for further benchmarking which can be useful for the evaluation of potential KGE models in the future. This provides a reasonable ground to compare our results against the existing English language datasets. Moreover, KGE models can be verified for our new dataset.

The rest of the paper is organized as follows: In section 2, we discuss related work about French KG and link prediction tasks. Section 3 explains our proposed methodology for building RezoJDM16k and also describes some of the state-of-the-arts KGE models used as our baselines. Section 4 explains our experimental setups and the parameters used for training KGE models with some discussions on the results. In the last section, we conclude our paper and discuss possible future works.

RezoJDM16k is freely available for public use[3].

2. Related Work

Current KGE models use extensively English KG datasets such as FB15k-237 and WN18RR (Toutanova et al., 2015). The models can be used for non-English KGs, nevertheless, there is no attempt for creation an efficient dataset for the French language usable for embedding models. There is RezoJDM15k[4], a dataset created by sub-selecting RezoJDM (Cousot et

[3]github.com/ContentSide/French_Knowledge_Graph

[4]The French KG dataset introduced in (Cousot et al., 2019) has no explicit name. We call it RezoJDM15k to discriminate it from our dataset in this paper.

al., 2019). It has 15K nodes, 43K triplets, 6 relation types and is used for link prediction tasks employing Random Forest Classifier. Since applying any classical machine learning algorithm demands feature engineering, Node2Vec approach (Grover and Leskovec, 2016) is used for converting nodes in RezoJDM15k to 20-dimension vectors. The Node2Vec approach can be categorized as a path-based model since it utilizes a second-order random walk approach to generate (sample) network neighborhoods for nodes. Path-based approaches have their own limitations : the larger the step size is, the larger the optimal solution space, but the computational complexity is higher (Wang et al., 2021).

RezoJDM15k has some technical limitations that can not efficiently be used in most of KGE algorithms: (i) the test dataset has a few common nodes with the training dataset, and it makes it difficult for modern KGE models to infer for instance a tail from a given head and relation as input; (ii) the number of relation types are 6 which is very limited from a practical point of view; (iii) there is no well-defined filtration criteria for KGE models since it was initially designed to be feature engineered from Node2Vec embedding and to be fed into Random Forest Classifier model. These are the limitations that we must evidently avoid in order to have more efficient KGE models.

The problem of path-based approaches (such as Node2Vec) can be treated by employing knowledge graph embedding (KGE) methods which have significantly advanced the state of the art. Knowledge graph embedding (KGE) or knowledge representation learning (KRL) is defined as learning a low-dimensional representations of a given knowledge graph. The low-dimensional embedding must preserve meanings of entities and relations in the original KG. KGE models are mainly used for missing link prediction task . According to (Wang et al., 2021) KGEs models can be categorized into three groups:

(i) The first category is translational-distance-based (or additive) models such as TransE (Bordes et al., 2013), TransH (Wang et al., 2014) and TransD (Ji et al., 2015). TransE regards the relations in KGs as translation vectors. Given a triplet (h, r, t), the relation r translates the head entity h to the tail entity t. It defines a scoring function (ψ) to measure the correctness of the triplet in the embedding space (e_o, r, e_s) as follows:

$$\psi(e_o, r, e_s) = -||h + r - t||_2^2$$

TransH defines a hyperplane for each relations, and translation property should be established on that hyperplane as follows:

$$h_\perp = w_r^\perp h w_r \ , \ t_\perp = w_r^\perp t w_r$$

$$\psi(e_o, r, e_s) = -||h_\perp + r - t_\perp||_2^2$$

TransD creates a dynamic matrix for all entity-relation pairs and maps the head and tail into M1 and M2, respectively. The transition from head to tail is as follow:

$$M_r^1 = w_r w_h^\perp + I \ , \ M_r^2 = w_r w_t^\perp + I$$
$$h_\perp = M_r^1 h \ , \ t_\perp = M_r^2 t$$
$$\psi(e_o, r, e_s) = -||h_\perp + r - t_\perp||_2^2$$

(ii) Semantic-matching-based (or multiplicative) models: DistMult (Yang et al., 2014) and ComplEx (Trouillon et al., 2016), which can outperform the additive models by capturing more semantic information. These models first embed entities and relations into a unified continuous vector space and then define a scoring function to measure its authenticity.

(iii) Neural-network-based models: such as ConvE (Dettmers et al., 2018) and SACN (Shang et al., 2019). These models consider the type of entity or relation, temporal information, path information and substructure information.

3. Proposed Methodology

In this section, we introduce our methodology for building the RezoJDM16k dataset. Firstly, we describe how the sub-selection on RezoJDM is performed. Consequently, we describe the performance indicators we used.

3.1. Graph Sub-Selection Algorithm

As we discussed in section 1, several sub-selection criteria are needed to make an efficient KG dataset (Socher et al., 2013; Bordes et al., 2013; Toutanova et al., 2015). These criteria are supposed to impact the quality and interpretability of the evaluations. For building our French KG dataset, we have taken into account some of these general criteria and guidelines. In addition, we have employed some specific filtering criteria that are only meaningful for RezoJDM lexical-semantic networks. Table 3 and the succeeding algorithm show in an abstract way, our input/output variables and our KG sub-selection algorithm, respectively:

Variable	Description
V_{in}	Set of nodes in RezoJDM
E_{in}	Set of edges in RezoJDM
r_{min}	Lowest frequency of relations
nd_{min}	Minimum node degrees
V_{out}	Set of nodes after filtering
E_{out}	Set of edges after filtering

Table 3: Inputs and outputs of the algorithm

Algorithm: Graph Sub-Selection Algorithm

GraphSubSelect *($V_{in}, E_{in}, r_{min}, nd_{min}$)*
- $V_1 \leftarrow$ FilterNodes(V_{in})
- $E_1 \leftarrow$ FilterEdges(E_{in})
- $V_2 \leftarrow$ UpdateNodes(V_1)
- $E_2 \leftarrow$ RemoveInverseRelations(E_1)
- $E_{out} \leftarrow$ FilterByRelationOccur(E_2, r_{min})
- $V_{out} \leftarrow$ FilterByNodesDegree(V_2, nd_{min})
- **return** (V_{out}, E_{out})

The algorithm operates on the original RezoJDM data. We note V_{in} and E_{in} the sets of nodes and edges, respectively. After applying a sequence of filters, the filtered graph, V_{out} and E_{out}, is obtained. Firstly, we filter out nodes and edges based on their types and weights. For nodes, only terms and their morphological variations are kept. Just a few relationship types are excluded, namely those related to chunks or internal implementation details irrelevant to our concerns. Both nodes and edges are filtered if their weight is lesser than 50. It is mandatory to apply $UpdateNodes$ to remove the nodes that has no edges after applying the $FilterEdges$. RezoJDM contains some relationship types that are symmetric, such as hypernymy/hyponymy or holonymy/meronymy. Such property allow the model to get the correct predictions by simply learning that a certain type is the inverse of another instead of actually modeling the relationship. Therefore the next essential step is applying $RemoveInverseRelations$ to prevent test leakage through inverse relations as described in (Toutanova et al., 2015). To do so, we look for the pattern $x \xrightarrow{t} y$ and $x \xleftarrow{t^{-1}} y$ and remove the edge with the lesser weights.

In order to make the Knowledge Graph more efficient for graph embedding models, we need to apply two more filters: $FilterByRelationOccur$ removes relationships with less than r_{min} occurrences; and $FilterByNodesDegree$ removes nodes with degrees less than nd_{min}. The experimental results show the near average choice of $r_{min} = 100$ and $nd_{min} = 45$ works more efficiently for building KGE models. We end up with 16k nodes, 832k triplets and 53 relation types. Following the common practice, we have named our dataset RezoJDM16k.

Finally, we splitted RezoJDM16k into three train, validation and test samples (90%, 5% and 5%). The statistics of RezoJDM16k are shown in Table 4. The comparison between RezoJDM16k with two popular English datasets are also available. WN18RR (Toutanova et al., 2015) is build from WordNet and is centered around hyponym/hyperonym relations. FB15k-237 (Dettmers et al., 2017) is based on Freebase.

Resource	Entities	triplets	Types
WN18RR (Train)	$41k$	$87k$	11
WN18RR (Validation)	$41k$	$3k$	11
WN18RR (Test)	$41k$	$3k$	11
FB15k-237 (Train)	$15k$	$272k$	237
FB15k-237 (Validation)	$15k$	$17k$	237
FB15k-237 (Test)	$15k$	$20k$	237
RezoJDM16k (Train)	$16k$	$666k$	53
RezoJDM16k (Validation)	$16k$	$83k$	53
RezoJDM16k (Test)	$16k$	$83k$	53

Table 4: Dataset statistics of the RezoJDM10k split compared to FB15k-237 and WN18RR splits

3.2. Performance Indicators

In the literature, there are three major metrics for measuring the quality of embedding models, namely, Hits@K, MR, and MRR (Chen et al., 2020). These metrics are frequently used and are fairly simple:

(i) *Hits@K*: is a performance index that measures the probability to find the correct prediction in the first top K model predictions (Chen et al., 2020). By convention K values varies between 1, 3, 5 and 10. The larger *Hits@K* values are, the better predictive performances.

(ii) *Mean Rank (MR)*: is the average ranking position of the items predicted by the model among all the possible items (Chen et al., 2020). The smaller the value, the better the model

(i) *Mean Reciprocal Rank (MRR)*: measures the number of triples predicted correctly (Chen et al., 2020). The larger the index, the better the model.

4. Experimental Setup

We used translational-based and semantic-based KGE modeling. In particular, we utilized TransE, TransH, TransD, DistMult and ComplEx for our KGE modelings. These models, unlike deep neural-network, are faster and computationally very efficient, as we discussed in section 2.

4.1. Baseline Graph Embedding Models

We tested the overall methodology described in section 3 to measure the performance on the RezoJDM16k dataset. We used the accuracy metrics MRR, MR, Hits@10, Hits@3 and Hits@1 to enable the head-to-head comparison with the state of the arts models and other reported baselines results. We chose our experimental hyper-parameters as follow: AdaGrad as optimizer (Duchi et al., 2011), alpha=0.5, batch size = 100 and number of epoch=200. The KGE models performance are illustrated in table 5.

Table 6 presents state-of-the-art performance on RezoJDM16k and the two English KG datasets (FB15K237

Model	MRR	MR	Hits@10	Hits@3	Hits@1
TransE	0.179	203.31	0.432	0.242	0.041
TransH	0.218	177.12	0.498	0.291	0.069
TransD	0.216	**170.68**	0.500	0.287	0.066
DistMult	0.220	194.47	0.445	0.252	0.109
ComplEx	**0.253**	201.58	**0.533**	**0.304**	**0.117**

Table 5: Performance of knowledge graph embedding models for RezoJDM16k

and WN18RR). To enable the head-to-head caparison with the English datasets, we used *Hits@10* evaluation metric as reported in the original papers for TransE (Bordes et al., 2013), TransH (Wang et al., 2014), TransD (Ji et al., 2015), DisMult (Yang et al., 2014) and ComplEx (Trouillon et al., 2016)

Model	RezoJDM16k	WN18RR	FB15k-237
TransE	0.432	0.501	0.486
TransH	0.498	0.507	**0.490**
TransD	0.500	0.508	0.487
DistMult	0.445	0.490	0.419
ComplEx	**0.533**	**0.510**	0.428

Table 6: Comparison of KGE state-of-the-models performance (Hits@10) between RezoJDM16k and English datasets

4.2. Discussions

Table 5 shows the performance of different KGE models on RezoJDM16k dataset using the evaluation metrics *MRR*, *MR*, *Hits@10*, *Hits@3* and *Hits@1*. ComplEx model has the best performance according to *MRR*, *Hits@10*, *Hits@3* and *Hits@1* indicators. TranseD shows better performance based on *MR* metric. One important observation is the superiority of semantic-based KGE models (DisMult and Comlex) over the translational-based KGE models (TransE, TranseH and TranseD) based on *MRR*, *MR*, and *Hits@1*. This is expected due to complexity of semantic-based models. Nevertheless, TransH and TransD provides competing performances based on *Hits@10* and *Hits@3* scores.

Table 6 shows the comparison of the performances of KGE state-of-the-art models trained on RezoJDM16k and two famous English KG datasets, namely, WN18RR and FB15k-237 using the evaluation metric *Hits@10*. In general, we observe that the performance scores of KGE models range from 0.428 till 0.528. This fact endorses that the learnability of RezoJDM16k is almost the same in terms of quantity, compared to WN18RR and FB15k-237. TransH, produce rather close scores for RezoJDM16k and FB15k-237. Whereas, ComplEx model shows close scores for RezoJDM16k and FB15k-237. For RezoJDM16k and WN18RR datasets the best performance score is for ComplEx model which have the highest number of triplets per relation type. For FB15k-237 dataset, the best performance belongs to TransH. To summarize, we can conclude that state of the arts KGE algorithms can learn the structure of KG presented in RezoJDM16k in an acceptable way.

5. Conclusions and Future Works

We introduced RezoJDM16k, a French Knowledge Graph dataset built from RezoJDM. The dataset consists of 16k nodes, 832k triplets with 53 different types in its train/dev/test datasets splits. We considered the incompleteness of the dataset as any KGs and addressed the task of link prediction to build a more complete KG. In this context, we provided a comparative study of strong predictive knowledge graph embedding models as baselines for future references. Furthermore, we compared the performance of these models with well-known English KGs, namely, FB15k-237 and WN18RR (Toutanova et al., 2015). The models exhibit similar performance of RezoJDM16k compared to other English KG datasets.

Many possible techniques could either enhance the quality of RezoJDM16k or empower the predictive capabilities with a more complex model. This includes, but is not limited to, enriching the current dataset with the polysemy that is encoded in RezoJDM. Additionally, we can employ the neural-network-based architectures introduced in ConvE (Dettmers et al., 2018) and SACN (Shang et al., 2019) which are not explored in this paper. One of the possible future studies is using/extending these architectures. Consequently, we can analyze the performance of our model against available French and English datasets.

The workflow presented in this paper, and availability of RezoJDM16k with KGE models, can pave the way for further directions of research in computational linguistics for French-based resources. To count some of them: (i) some symbolic approaches (Lafourcade et al., 2018) use RezoJDM to create lexicons for type-theoretic frameworks for compositional semantics. Our introduced approach can augment this sort of study with the prediction of complex linguistic type-shifting cases that do not explicitly exist in RezoJDM; (ii) enhancing the quality of lexical-quantifier preference problem (Catta and Mirzapour, 2017) with feeding the systems with semantic relations between the headwords in multiple-quantifiers ambiguous sentences; (iii) there are some studies for measuring linguistic complexity on syntactic level (Zou et al., 2022; Mirzapour et al., 2020) that use dependency-like relations between words in a sentence. The syntactic nature of relations can be augmented with semantic relations. This provides a rich formalism for psycholinguistic theories that use semantic relations on the word level.

6. Bibliographical References

Arora, S. (2020). A survey on graph neural networks for knowledge graph completion. *arXiv preprint arXiv:2007.12374*.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

Catta, D. and Mirzapour, M. (2017). Quantifier Scoping and Semantic Preferences. In *CONLI: Computing Natural Language Inference*, Montpellier, France.

Chatzikyriakidis, S., Lafourcade, M., Ramadier, L., and Zarrouk, M. (2017). Type theories and lexical networks: Using serious games as the basis for multi-sorted typed systems. *Journal of Language Modelling*, 5(2):229–272.

Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., and Duan, Z. (2020). Knowledge graph completion: A review. *IEEE Access*, 8:192435–192456.

Cousot, K., Mirzapour, M., and Ragheb, W. (2019). Prediction of missing semantic relations in lexical-semantic network using random forest classifier. *arXiv preprint arXiv:1911.04759*.

Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2017). Convolutional 2D Knowledge Graph Embeddings. *CoRR*, abs/1707.0.

Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

Ji, G., He, S., Xu, L., Liu, K., and Zhao, J. (2015). Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696.

Lafourcade, M. and Le Brun, N. (2017). Extracting semantic relations via the combination of inferences, schemas and cooccurrences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 417–423.

Lafourcade, M., Mery, B., Mirzapour, M., Moot, R., and Retoré, C. (2018). Collecting weighted coercions from crowd-sourced lexical data for compositional semantic analysis. In Sachiyo Arai, et al., editors, *New Frontiers in Artificial Intelligence*, pages 214–230, Cham. Springer International Publishing.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand.

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.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mirzapour, M., Prost, J.-P., and Retoré, C., (2020). *Measuring Linguistic Complexity: Introducing a New Categorial Metric*, pages 95–123. Springer International Publishing, Cham.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Shang, C., Tang, Y., Huang, J., Bi, J., He, X., and Zhou, B. (2019). End-to-end structure-aware convolutional networks for knowledge base completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3060–3067.

Socher, R., Chen, D., Manning, C. D., and Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge.

Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., and Gamon, M. (2015). Representing Text for Joint Embedding of Text and Knowledge Bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal, sep. Association for Computational Linguistics.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR.

Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). Knowledge graph embedding by translating on hy-

perplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Wang, M., Qiu, L., and Wang, X. (2021). A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3):485.

Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zou, L., Carl, M., Mirzapour, M., Jacquenet, H., and Vieira, L. N. (2022). Ai-based syntactic complexity metrics and sight interpreting performance. In Jong-Hoon Kim, et al., editors, *Intelligent Human Computer Interaction*, pages 534–547, Cham. Springer International Publishing.