French Lexical Semantic Graphs: Enrichment by Link Prediction and Integration in a WSD model

Hee-Soo Choi

ANR SELEXINI Meeting

Quick presentation

- 3rd year PhD student in Language Science at ATILF and LORIA, Nancy
- Supervised by Mathieu Constant, Karën Fort and Bruno Guillaume
- MSc and BSc in Linguistics and Computer Science in Sorbonne University, Paris

Main themes of the PhD

- Overview on French lexical resources [Choi, 2022, Choi et al., 2023]
- Enriching French lexical semantic graphs with link prediction [Choi et al., 2024]
- Leveraging linguistic information in graph embeddings
- Improving Word Sense Disambiguation task in French

Beyond Model Performance: Can Link Prediction Enrich French Lexical Graphs?

Hee-Soo Choi, Priyansh Trivedi,
Mathieu Constant, Karën Fort, Bruno Guillaume

LREC-COLING 2024, Turin, Italy

Motivations

Context:

- Knowledge graphs and lexical graphs are incomplete

Motivations

Context:

- Knowledge graphs and lexical graphs are incomplete
- Link Prediction task addresses this issue but mostly focuses on model performance

Motivations

Context:

- Knowledge graphs and lexical graphs are incomplete
- Link Prediction task addresses this issue but mostly focuses on model performance
- Most of the work conducted on English language

Motivations

Context:

- Knowledge graphs and lexical graphs are incomplete
- Link Prediction task addresses this issue but mostly focuses on model performance
- Most of the work conducted on English language

We propose:

- a resource-oriented approach on two French lexical graphs

Motivations

Context:

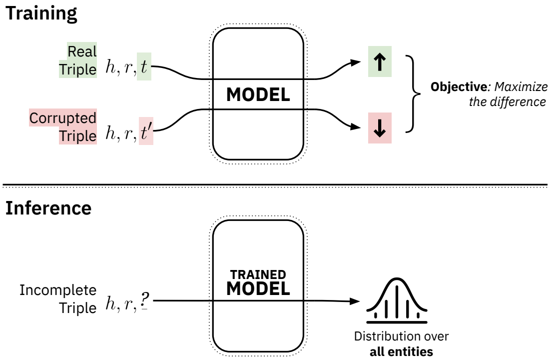
- Knowledge graphs and lexical graphs are incomplete
- Link Prediction task addresses this issue but mostly focuses on model performance
- Most of the work conducted on English language

We propose:

- a resource-oriented approach on two French lexical graphs
- to extract new relations from a link prediction model to enrich a sparse lexical graph

Link prediction task

The link prediction task consists in predicting missing triples in a graph described by a set of triples (h, r, t) for head, relation and tail.

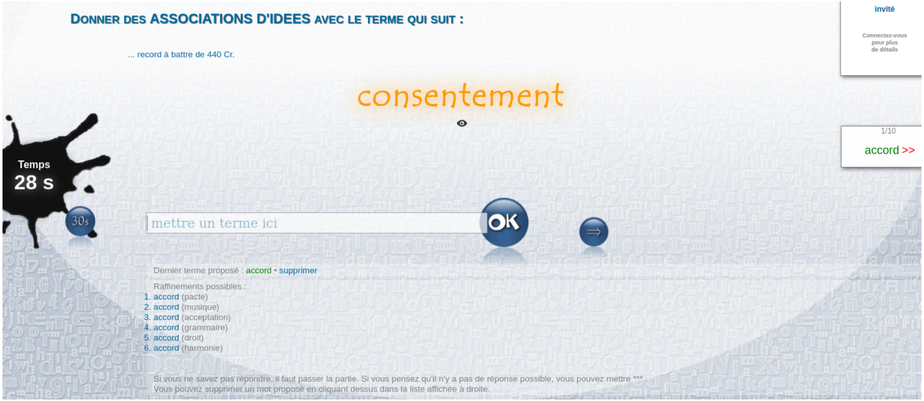


REZO and JeuxDeMots [Lafourcade and Joubert, 2008]

- Very dense resource: 6 million nodes and 537 million edges in October 2023
- Made with GWAPs and semi-automatic mechanisms

REZO and JeuxDeMots [Lafourcade and Joubert, 2008]

- Very dense resource: 6 million nodes and 537 million edges in October 2023
- Made with GWAPs and semi-automatic mechanisms

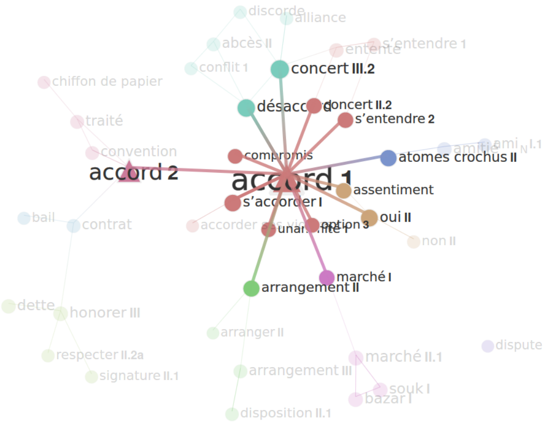


Réseau lexical du français (RL-fr) [Lux-Pogodalla and Polguère, 2011]

- 29,220 nodes and 72,054 edges
- Created manually and based on the Meaning-Text Theory [Mel'čuk, 1996]

Réseau lexical du français (RL-fr) [Lux-Pogodalla and Polguère, 2011]

- 29,220 nodes and 72,054 edges
- Created manually and based on the Meaning-Text Theory [Mel'čuk, 1996]



Datasets for French Link Prediction

- `RezoJDM16k` [Mirzapour et al., 2022] and `RLF27k`
- Transductive Link Prediction configuration
- Division into 80%, 10%, 10%

	`RezoJDM16k`	`RLF27k`
# nodes	15,746	27,068
# edges	832,093	71,017
# triples Train	665,674	57,643
# triples Valid	83,209	6,674
# triples Test	83,210	6,700

Metrics based on predictions' scores

- MR (Mean Rank): average rank of the positive triples
- MRR (Mean Reciprocal Rank): average of the reciprocal of ranks of the positive triples
- Hits@k: proportion of positive triples in the top k ranked triples

Training Link Prediction models on RezoJDM16k and RLF27k

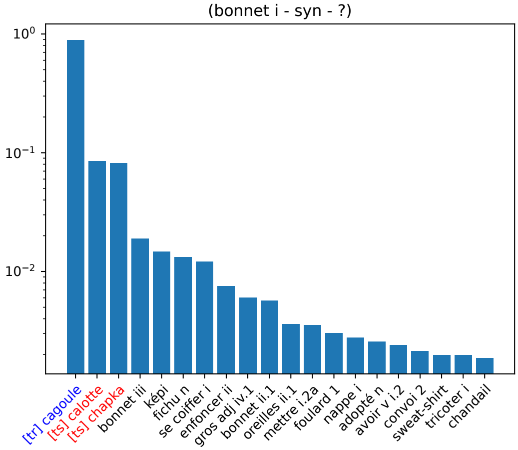
Model (For RezoJDM16k)	**MRR ↑**	**MR ↓**	**Hits@10 ↑**	**Hits@3 ↑**	**Hits@1 ↑**
TransE [Bordes et al., 2013]	0.180	200.78	0.437	0.242	0.040
TransH [Wang et al., 2014]	0.217	173.28	0.503	0.293	0.064
TransD [Ji et al., 2015]	0.216	**168.18**	0.500	0.290	0.065
DistMult [Yang et al., 2015]	0.219	194.16	0.446	0.252	0.109
ComplEx [Trouillon et al., 2016]	0.256	190.79	0.539	0.309	0.119
RotatE [Sun et al., 2019]	0.312	177.04	0.587	0.409	0.155
CompGCN-ConvE [Vashishth et al., 2020]	**0.461**	171.26	**0.659**	**0.514**	**0.357**

Training Link Prediction models on RezoJDM16k and RLF27k

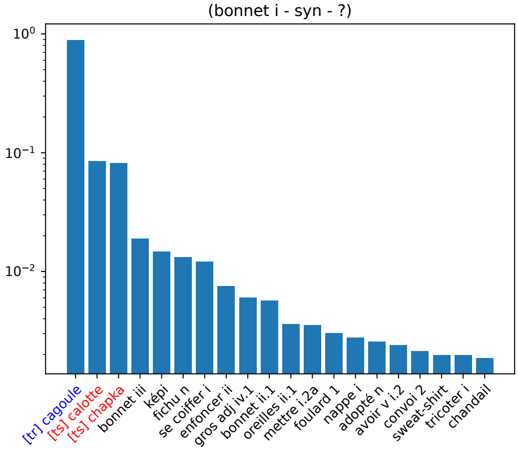
Model (For RezoJDM16k)	**MRR ↑**	**MR ↓**	**Hits@10 ↑**	**Hits@3 ↑**	**Hits@1 ↑**
TransE [Bordes et al., 2013]	0.180	200.78	0.437	0.242	0.040
TransH [Wang et al., 2014]	0.217	173.28	0.503	0.293	0.064
TransD [Ji et al., 2015]	0.216	**168.18**	0.500	0.290	0.065
DistMult [Yang et al., 2015]	0.219	194.16	0.446	0.252	0.109
ComplEx [Trouillon et al., 2016]	0.256	190.79	0.539	0.309	0.119
RotatE [Sun et al., 2019]	0.312	177.04	0.587	0.409	0.155
CompGCN-ConvE [Vashishth et al., 2020]	**0.461**	171.26	**0.659**	**0.514**	**0.357**

Model (For RLF27k)	**MRR ↑**	**MR ↓**	**Hits@10 ↑**	**Hits@3 ↑**	**Hits@1 ↑**
TransE [Bordes et al., 2013]	0.278	**2594.24**	0.624	0.497	0.033
TransH [Wang et al., 2014]	0.250	2957.59	0.581	0.465	0.011
TransD [Ji et al., 2015]	0.255	2752.03	0.587	0.472	0.016
DistMult [Yang et al., 2015]	0.373	2748.25	0.613	0.502	0.216
ComplEx [Trouillon et al., 2016]	0.413	3447.98	0.593	0.524	0.284
RotatE [Sun et al., 2019]	0.399	3650.92	0.490	0.454	0.336
CompGCN-ConvE [Vashishth et al., 2020]	**0.515**	2808.68	**0.627**	**0.559**	**0.450**

Analyzing CompGCN-ConvE model's predictions

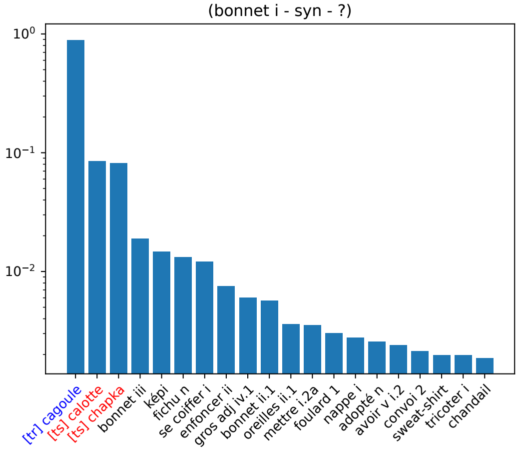


Analyzing CompGCN-ConvE model's predictions



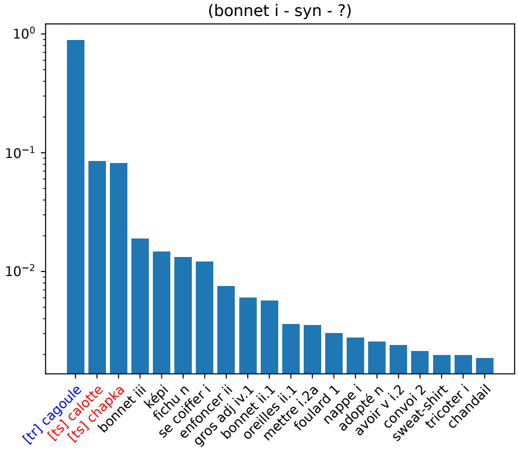
- [*tr*] cagoule - 0.893

Analyzing CompGCN-ConvE model's predictions



- [*tr*] cagoule - 0.893
- [*ts*] calotte - 0.085
- [*ts*] chapka - 0.082

Analyzing CompGCN-ConvE model's predictions

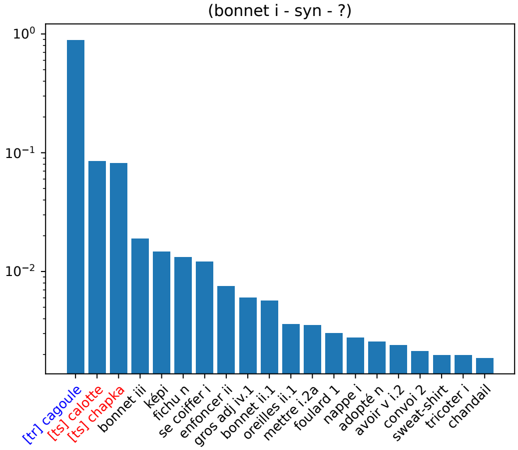


- [*tr*] cagoule - 0.893

- [*ts*] calotte - 0.085
- [*ts*] chapka - 0.082

- Triples not in the graph: < 0.02

Analyzing CompGCN-ConvE model's predictions



- [*tr*] cagoule - 0.893
- [*ts*] calotte - 0.085
- [*ts*] chapka - 0.082
- Triples not in the graph: < 0.02

$\rightarrow$ Function score only can't discriminate relevant new triples

Computing a confidence score with Monte Carlo Dropout

During inference, we apply Monte Carlo Dropout [Gal and Ghahramani, 2016] :

Computing a confidence score with Monte Carlo Dropout

During inference, we apply Monte Carlo Dropout [Gal and Ghahramani, 2016] :

- Dropout: **Randomly switching off neurons** in a neural network

Computing a confidence score with Monte Carlo Dropout

During inference, we apply Monte Carlo Dropout [Gal and Ghahramani, 2016] :

- Dropout: **Randomly switching off neurons** in a neural network
- **100 output distributions** for the same input by sampling **different dropout mask**

Computing a confidence score with Monte Carlo Dropout

During inference, we apply Monte Carlo Dropout [Gal and Ghahramani, 2016] :

- Dropout: **Randomly switching off neurons** in a neural network
- **100 output distributions** for the same input by sampling **different dropout mask**
- We compute **how many times a prediction appears in the top 10**.
  Example: If it appears 60 times in the top 10, the confidence score is 60%.

Extracting candidates triples

- We compute the confidence score for all possible combinations of triples for `RezoJDM16k` and `RLF27k`.

Extracting candidates triples

- We compute the confidence score for all possible combinations of triples for `RezoJDM16k` and `RLF27k`.
- Triples already existing in the graphs are removed.

Extracting candidates triples

- We compute the confidence score for all possible combinations of triples for `RezoJDM16k` and `RLF27k`.
- Triples already existing in the graphs are removed.
- For `RLF27K`, we extract **triples whose entities are not linked by an oriented path** in the graph: 95,766 triples.

Extracting candidates triples

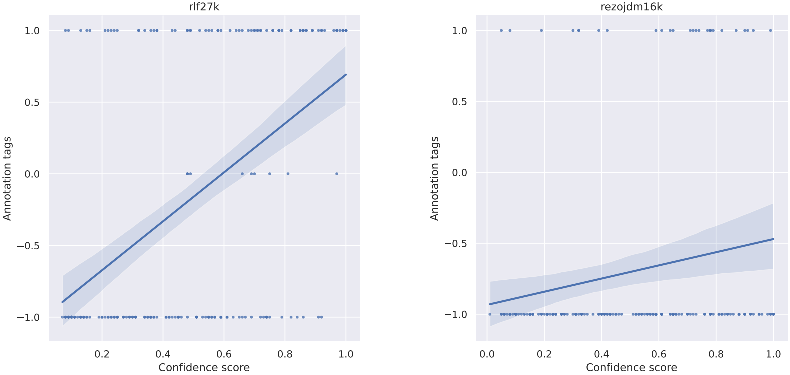
- We compute the confidence score for all possible combinations of triples for `RezoJDM16k` and `RLF27k`.
- Triples already existing in the graphs are removed.
- For `RLF27K`, we extract **triples whose entities are not linked by an oriented path** in the graph: 95,766 triples.
- For `RezoJDM16k`, we extract **triples whose entities are furthest apart** (maximum path size 3 and 4): 154,168 triples.

Evaluating confidence score with manual annotations

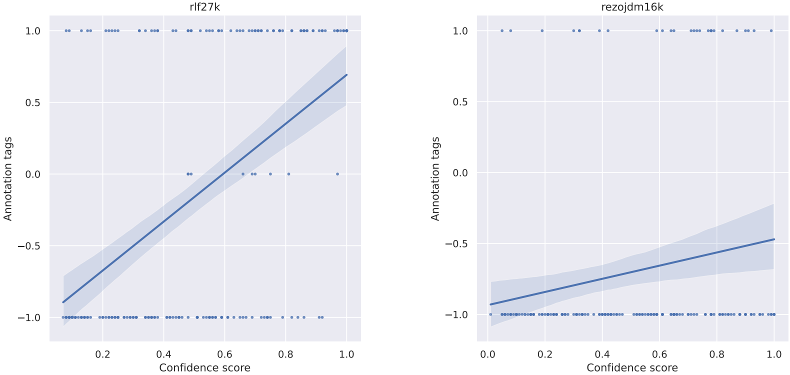
Annotation of 240 triples by 4 annotators for each dataset.
The task is to determine if two entities are linked with semantic or syntactic relation.
Three annotation tags are used:

- 1: there is a link between the entities
- -1: there is no link
- 0: the link is ambiguous or questionable

Correlation between annotations and confidence scores

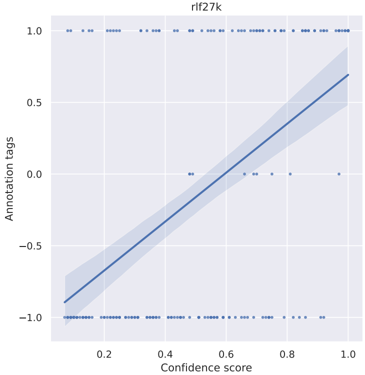


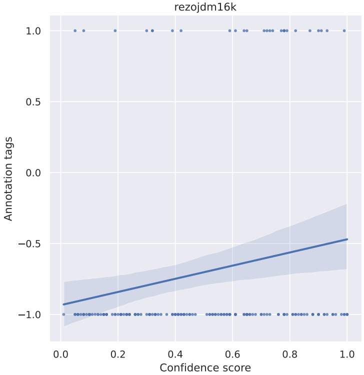
Correlation between annotations and confidence scores



- RLF27k: high correlation - triples with high confidence score are relevant

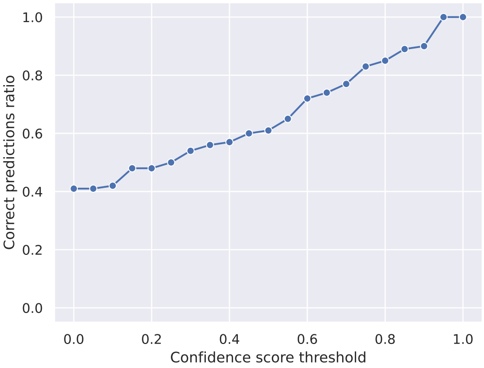
Correlation between annotations and confidence scores





- `RLF27k`: high correlation - triples with high confidence score are relevant
- `RezoJDM16k`: poor correlation due to high density of the graph, two nodes semantically different are related with a relatively short path

Determining a confidence score threshold for RLF27k



→ A confidence threshold of 0.95 results in 100% of triples annotated as correct in RLF27k, which gives us **398 potential good triples** out of the 95,766 candidates.

Relevant new triples for `RLF27k`

- `(kidnappeur, Syn, ravisseur I)` *(kidnapper, Syn, abductor I)*
- `(marchande, Syn, débitante)` *(merchant, Syn, retailer)*
- `(motocycliste n-fem, Syn, motarde)` *(motorcyclist n-fem, Syn, biker)*

Refined triples in RezoJDM16k

- 31% of the edges in RezoJDM16k are the general relation associated

In RezoJDM16k	In CompGCN-ConVE's predictions
(infirmière, associated, personne)	(infirmière, is_a, personne)
(herpès, associated, médecine)	(herpès, domain, médecine)
(ouvrir, associated, fermer)	(ouvrir, antonym, fermer)

Conclusion

Contributions:

- Link prediction on 2 French lexical semantic graphs with 7 models
- Addition of a confidence score to CompGCN-ConvE model's predictions
- Qualitative analysis of predictions based on manual annotations
- Extraction of new triples in `RL-fr`

Conclusion

Contributions:

- Link prediction on 2 French lexical semantic graphs with 7 models
- Addition of a confidence score to CompGCN-ConvE model's predictions
- Qualitative analysis of predictions based on manual annotations
- Extraction of new triples in `RL-fr`

Limitations:

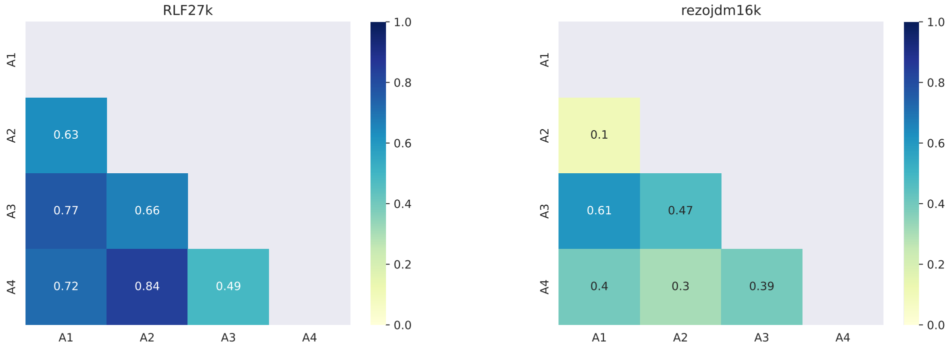
- Need for manual verification of candidate triples
- Influence of the representation of polysemy in different nodes in the `RL-fr`

Work in progress...

- Integrating graph embeddings trained in Link Prediction into EWISER model [Bevilacqua and Navigli, 2020]
- Testing on RL-fr lexicographical examples [Sinha et al., 2022]
- Leveraging supersenses to generate semi-automatically WSD annotations on `fr-SemCor` [Barque et al., 2020] with AMuSE-WSD [Orlando et al., 2021]

Thank you for your attention
Questions?

Inter-annotators agreements



Barque, L., Haas, P., Huyghe, R., Tribout, D., Candito, M., Crabbé, B., and Segonne, V. (2020).
FrSemCor: Annotating a French corpus with supersenses.
In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 5912–5918, Marseille, France. European Language Resources Association.

Bevilacqua, M. and Navigli, R. (2020).
Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information.
In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2854–2864, Online. Association for Computational Linguistics.

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., and Yakhnenko, O. (2013).
Translating embeddings for modeling multi-relational data.

In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, page 2787–2795, Red Hook, NY, USA. Curran Associates Inc.

Choi, H.-S. (2022).
État de l'art : Liage de ressources lexicales du français.
In
RÉCITAL 2022 - 24e Rencontre des Étudiants Chercheurs en Informatique pour le Traiten
Avignon, France.

Choi, H.-S., Fort, K., Guillaume, B., and Constant, M. (2023).
Des ressources lexicales du français et de leur utilisation en TAL : étude des actes de TALN.
In
TALN 2023 - Conférence sur le Traitement Automatique des Langues Naturelles,
Paris, France.

Choi, H.-S., Trivedi, P., Constant, M., Fort, K., and Guillaume, B. (2024).
Beyond model performance: Can link prediction enrich French lexical graphs?

In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2329–2341, Turin, Italie. ELRA and ICCL.

Gal, Y. and Ghahramani, Z. (2016).
Dropout as a bayesian approximation: Representing model uncertainty in deep learning.
In Balcan, M. F. and Weinberger, K. Q., editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 1050–1059, New York, New York, USA. PMLR.

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, Beijing, China. Association for Computational Linguistics.

Lafourcade, M. and Joubert, A. (2008).

JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes.
In
JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles,
pages 657–666, France.

Lux-Pogodalla, V. and Polguère, A. (2011).
Construction of a French Lexical Network: Methodological Issues.
In First InternationalWorkshop on Lexical Resources, WoLeR 2011, pages 54–61, Ljubljana, Slovenia.

Mel'čuk, I. (1996).
Lexical functions in lexicography and natural language processing.
Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon,
pages 37–102.

Mirzapour, M., Ragheb, W., Saeedizade, M. J., Cousot, K., Jacquenet, H., Carbon, L., and Lafourcade, M. (2022).
Introducing RezoJDM16k: a French KnowledgeGraph DataSet for link prediction.
In Proceedings of the Thirteenth Language Resources and Evaluation Conference,
pages 5163–5169, Marseille, France. European Language Resources Association.

Orlando, R., Conia, S., Brignone, F., Cecconi, F., and Navigli, R. (2021).
AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation.
In Adel, H. and Shi, S., editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 298–307, Online et Punta Cana, République dominicaine. Association for Computational Linguistics.

Sinha, A., Ollinger, S., and Constant, M. (2022).
Word sense disambiguation of French lexicographical examples using lexical networks.
In Ustalov, D., Gao, Y., Panchenko, A., Valentino, M., Thayaparan, M., Nguyen, T. H., Penn, G., Ramesh, A., and Jana, A., editors, Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, pages 70–76, Gyeongju, Corée du Sud. Association for Computational Linguistics.

Sun, Z., Deng, Z., Nie, J., and Tang, J. (2019).
Rotate: Knowledge graph embedding by relational rotation in complex space.

In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., and Bouchard, G. (2016).
Complex embeddings for simple link prediction.
In Balcan, M. F. and Weinberger, K. Q., editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 2071–2080, New York, New York, USA. PMLR.

Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. P. (2020).
Composition-based multi-relational graph convolutional networks.
In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014).
Knowledge graph embedding by translating on hyperplanes.
Proceedings of the AAAI Conference on Artificial Intelligence, 28.

Yang, B., Yih, W., He, X., Gao, J., and Deng, L. (2015).
Embedding entities and relations for learning and inference in knowledge bases.

In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.