Word Sense Disambiguation of French Lexicographical Examples Using Lexical Networks

Aman Sinha, Sandrine Ollinger, Mathieu Constant

Word Sense Disambiguation of French Lexicographical Examples Using Lexical Networks

Aman Sinha, Sandrine Ollinger, Mathieu Constant
ATILF, Université de Lorraine, Nancy, France
{firstname.lastname}@atilf.fr

Abstract

This paper focuses on the task of word sense disambiguation (WSD) on lexicographic examples relying on the French Lexical Network (*fr-LN*). For this purpose, we exploit the lexical and relational properties of the network, that we integrated in a feedforward neural WSD model on top of pretrained French BERT embeddings. We provide a comparative study with various models and further show the impact of our approach regarding polysemic units.

1 Introduction

Word sense disambiguation is a long-standing research field in NLP investigating supervised, unsupervised, knowledge-based and mixed approaches (Navigli, 2009). Lexical resources have always played a crucial role not only serving as sense inventories, but also as sources of information to help the disambiguation process (a.o. Wilks and Stevenson (1998)). In particular, the structure and lexical content of lexical networks have been successfully exploited for this task with graph-based algorithms (a.o. Agirre et al. (2006)).

With the deep learning revolution, supervised approaches relying on neural networks and pretrained word embeddings have quickly gained popularity. In such framework, WSD is often seen as a token classification task, where tokens are assigned a sense label among an exist set of senses. Classical supervised models are built on a MultiLayer Perceptron (MLP) for predicting a sense label for the target tokens (Raganato et al., 2017) and lately the use of pretrained contextualized word embedding has become standard (ex. Vial et al. (2019)).

Such supervised systems are dependent on sense-annotated datasets that tend to have limited coverage due to the manual annotation cost. Furthermore, in these systems, rare senses are often disadvantaged towards more frequent ones. To tackle this problem, more and more research works propose approaches integrating lexical network knowledge to such models. Several strategies have been proposed: either integrating lexical knowledge – e.g. glosses (Huang et al., 2019) –, or integrating structural properties – e.g. use of graph-based algorithms such as Personalized PageRank (El Sheikh et al., 2021), use of hyperonym/hyponym/synonym relations in a lexical network to compress the sense tagset and then make the labeling task easier (Vial et al., 2019) –. Other models such as EWISE (Kumar et al., 2019) and EWISER (Bevilacqua and Navigli, 2020) enhance the WSD system with explicit and implicit knowledge using graph structure information from lexical knowledge networks and existing sense embeddings.

In this paper, we are interested in adapting the EWISER model to specific lexical data: the data from the French Lexical Network (*fr-LN*, Polguère (2014)) and its derived database of lexicographical usage examples (DBLE-LN-fr). In particular, we exploited the linguistic richness of its relation types, by integrating trainable weighted relations. Our system gets better or comparable results than the original system.

This paper is organized as follows. Section 2 presents our dataset and its particularities. Section 3 introduces the model and its adaptations. Sections 4 and 5 are respectively devoted to introducing the experimental setup and discussing and comparing the results.

2 The French Lexical Network and its database of lexicographical examples

2.1 A linguistically-rich lexical network

Lexical networks used as lexical knowledge in NLP are generally variants of WordNet (Miller, 1995). In this paper, we rely on the French lexical network *fr-LN*[1], which is under construction. It is based on the model of lexical systems (Polguère,

[1]The data are available on the ORTOLANG platform: https://hdl.handle.net/11403/lexical-system-fr/v2.1

2014) and is in line with the research projects conducted in the framework of Explanatory and Combinatorial Lexicology (Mel'čuk, 2006). It contains among others syntagmatic, paradigmatic, copolysemic and phraseological relations. The complete *fr-LN* contains 29,220 word senses and 80,036 relations between them. In this paper, we focus only on the 62,641 paradigmatic and syntagmatic links, which are standardized using the system of 686 distinct Meaning-Text lexical functions (LFs) (Polguère, 2007). Table 1 shows statistics on *fr-LN*. It differs from WordNet (WN) in several dimensions: WN has much larger coverage, contains few relation types that are mainly paradigmatic relations and is built on synset nodes. fr-LN relations mainly involve senses of different part-of-speech tags, whereas WN relations quasi-exclusively involve nodes of the same part-of-speech. For instance, less than 6% of the relations involving verbs are between two verbs. WN and fr-LN have comparable polysemy rates. Contrary to WN, fr-LN does not include glosses and the lexicographic definitions are still prototypical. An interesting feature of fr-LN is that relations are associated manually crafted semantic weights (three possible values: 0, 1 and 2) depending to what extent the semantic content of the source node includes the semantic content of the target one.

Graph	#Word Senses	#Lemmas	#LF-Arcs	#LFs
Complete	29,220	18,400	62,641	686
Verbs-only	5,237	2,559	9,854	399
Nouns-only	14,044	8,639	21,580	501

Table 1: Statistics on the *fr-LN* network.

2.2 The *DBLE-LN-Fr* database of lexicographical examples

The fr-LN lexical network comes with lexicographical usage examples for each word sense, that have been gathered in the *DBLE-LN-Fr* database[2]. The examples come from three main sources: *Frantext*[3], *FrWaC* (Baroni et al., 2009), the *Est-Républicain* newspaper corpus (ATILF and CLLE, 2020). They have been selected because they display interesting use cases for distinguishing meanings. They should enable speakers to appropriate the lexicographic descriptions of the lexical units they illustrate. Coupled with these descriptions, they provide all the information needed to use correctly each lexical unit described.

Corpus	#examples	#targets	#Word Senses	#Lemmas
Complete	31,131	51,347	27,343	17,161
Verbs-only	8,169	9,428	5,141	2,483
Nouns-only	19,644	27,105	13,601	8,131

Table 2: DBLE-LN-fr dataset. # targets stands for the number of occurrences of target words in the dataset.

Each example contains from one to eleven occurrences of lexical entities present in the fr-LN. These occurrences are marked and associated with the part-of-speech tag of the lexical entity and a link to visualize the lexical entity in the spiderlex web application[4]. For this work, we selected the examples which contain an occurrence of verb/noun word senses, excluding the examples that contain an occurrence of a verb/noun that is itself included in an occurrence of a multiword unit (locution, idioms, etc.). The table (2) synthesizes the composition of the resulting corpora. Figure 2 (resp. Figure 3) represents a subgraph for the lemma `ping-pong` from the lexical network fr-LN with all lexical function relations (resp. with relations with nouns only).

3 A model integrating graph knowledge

The proposed model is a variant of EWISER (Bevilacqua and Navigli, 2020) that we adapted using some specific features of fr-LN, namely the richness of its relation types, and the semantic weights associated to relations (cf. section 2). EWISER can be seen as a token classification system. It takes as input a sequence of words that feeds a BERT layer. For each target word, a feedforward module is then applied to predict its sense label given the input sequence. The exact modelling is depicted by the equation 1 taken and derived directly from the original paper[5].

$$\begin{aligned} H_0 &= \texttt{BatchNorm}(B) \\ H_1 &= \texttt{swish}(H_0 W + b) \\ Q &= H_1 A^T + H_1 \end{aligned} \tag{1}$$

In the above equation, B corresponds to the sum of the last four *BERT* hidden layer, which

[2]The data are available on the ORTOLANG platform: https://hdl.handle.net/11403/examples-ls-fr/v2

[3]https://www.frantext.fr

[4]https://spiderlex.atilf.fr/

[5]We removed from the original paper the use of external preexisting semantic embeddings as our aim was to rely entirely on the database of lexicographic examples and the French lexical network to evaluate their impact on the WSD task.

is given as input to a 2-layer feedforward to compute the logits H_1. This output is encoded with graph information from the lexical network using `A` which is the corresponding adjacency matrix. Each node corresponds to a possible word sense in the training dataset. In the original EWISER paper, matrix `A` encodes hypernym and hyponyms relations from Wordnet, whereas in our case it encodes paradigmatic and syntagmatic relations from fr-LN. The parameters of A may be frozen or trainable (Bevilacqua and Navigli, 2020).

In this paper, we use two strategies to compute the elements $a_{i,j}$ of A relying on some features of *fr-LN*. Every node pair (i, j) have a set $S_{i,j}$ of present relations between i and j. Each relation r has a weight $w(r)$, and $a_{i,j}$ is the sum of the weights of the relations between i and j: $a_{i,j} = \sum_{r \in S_{i,j}} w(r)$.

We consider two weighing schemes for every relation r: (1) $w(r) = 1$, the element $a_{i,j}$ being the cardinality of $S_{i,j}$ [STRUCT]; (2) $w(r) = s_r + 1$ where $s_r \in 0, 1, 2$ is the semantic weight of r, $a_{i,j}$ determining to what extent the semantic content of i is included in the one of j [SEM]. The STRUCT strategy is taken from (Bevilacqua and Navigli, 2020), whereas SEM is a contribution of this paper.

For each weighting scheme, we experimented three settings: (a) the element $a_{i,j}$ is frozen, (b) $a_{i,j}$ is trainable, (c) $w(r)$ is trainable, the weight of each relation being learnt from the training dataset. The setting (c) is a proposal of this paper, whereas (a) and (b) are taken from (Bevilacqua and Navigli, 2020).

4 Experimental Setup

4.1 Dataset

As stated in section 2, we experiment our models on the database of lexicographic examples DBLE-LN-fr built on the French lexical network Fr-LN (Polguère, 2014) focusing on nouns and verbs.

We performed a strategy-based data splitting using the following rules :

1. If the lemma has only one sense, we keep it in the train set, in order to prevent from having straightforward cases in the evaluation;
2. All lemma in test/dev should be in train;
3. Unseen senses can be in test/dev;
4. The distribution of senses between train and test/dev is proportional;
5. Any example with multiple senses to disambiguate should be in the same data split.

4.2 Baselines

We compare our variants of EWISER with various standard baselines. These include Most/Least Frequent Sense per lemma (MFS/LFS) baseline; a random sense (RS) baseline; a cosine-based similarity of the sense representations from BERT-based language model as (Barycenter) baseline (Le et al., 2020) and H_1 representation (refer eqn 1) as MLP baseline.

4.3 Implementation

We used contextual embeddings of two French language models namely, FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020). We use hidden layer size of 3000 and 8000 by rough estimate of number of unique lemmas in the verb and noun corpora respectively. We use `Adam` optimizer with learning rate 0.001 as a common setting for both sets of experiments. We use negative log likelihood (NLL) as our loss function. For each experiment, we used the following decoding strategy selecting the most probable sense among the possible senses for the target word in the fr-LN sense inventory. The code of this implementation is available on GitHub (`https://github.com/ATILF-UMR7118/GraphWSD`).

System	VERB		NOUN	
	Dev	Test	Dev	Test
MFS	0.1145	0.1427	0.2026	0.2016
LFS	0.1178	0.1091	0.1973	0.1939
RS	0.1578	0.1654	0.2444	0.2357
BARYC.	0.3189	0.3178	0.5390	0.5454
MLP	0.2648	0.2822	0.5091	0.5163
STRUCT	0.3513	0.3751	0.5061	0.5171
STRUCT*	0.3502	0.3708	**0.5521**	**0.5615**
STRUCT**	0.3372	0.347	0.5444	0.5516
SEM	0.3416	0.3676	0.5260	0.5309
SEM*	0.3556	0.3546	0.5379	0.5362
SEM**	**0.3610**	**0.3838**	0.5103	0.5274

Table 3: WSD results on DBLE-LN-fr. STRUCT and SEM are the two strategies to compute A matrix. By default, $a_{i,j}$ are frozen. * indicates that $a_{i,j}$ is trainable. ** indicates that the relation weights w are trainable.

5 Results and discussion

To evaluate our models, we used the accuracy of the system predictions, i.e. the percentage of correct

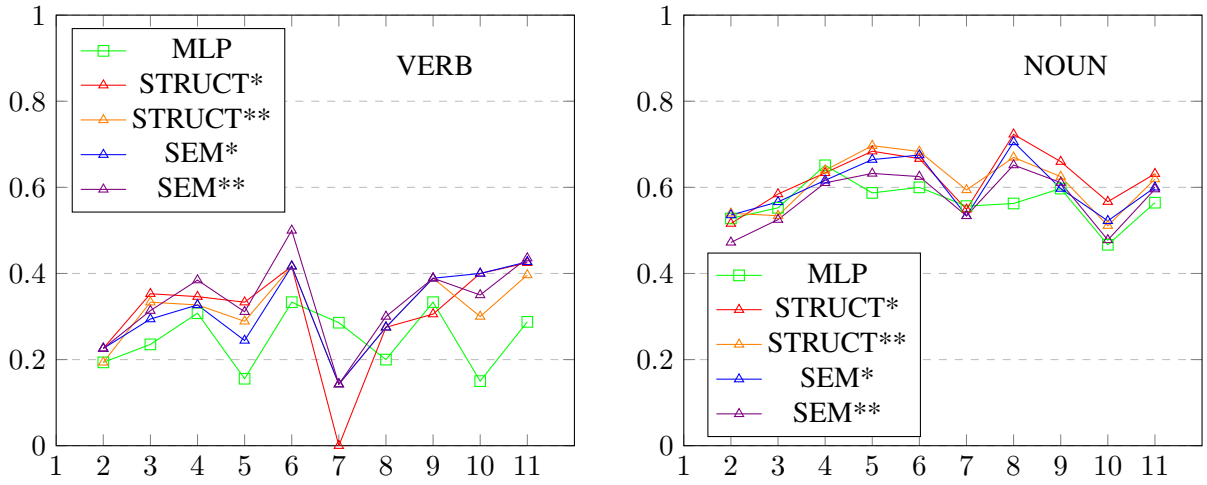

Figure 1: Polysemic performance analysis on dev set; x-axis: `sense-count` and y-axis : `accuracy`

predictions. The system was preliminary tuned on the dev dataset. The MLP-baseline obtained better performances using the CamemBERT embeddings, whereas the Barycenter performances were better using FlauBERT.

5.1 Global system performances

Table 3 shows results on both dev and test sets for all experimented systems both for nouns and verbs. Results are consistent across test and dev sets. MFS/LFS baselines results are on par with the random baseline, due to the uniform distribution of senses in our dataset coming from the use of lexicographic examples instead of standard annotated texts on which MFS is traditionally quite high. It is also worth noting that the simple Barycenter baseline consistently outperforms the MLP baseline. Our experiments consolidate the results of Bevilacqua and Navigli (2020), showing the integration of lexical network knowledge systematically tends to improve the WSD performances. Regarding the two strategies to compute the A matrix, SEM weights tend to perform better than STRUCT weights for verbs, whereas this is the other way around for nouns. In both cases, the use of trainable weights is favourable. The better performance of SEM for verbs can be attributed to the #LF-Arcs – #Lemma ratio (refer Table:1) which is more for verbs (3.85) than nouns (2.49) implying the semantic richness of the verb subgraph.

Overall, WSD on our dataset for French verbs is harder than for nouns (1/3 vs. 1/2 accuracy). We compared these results using those obtained for other French datasets. In particular, we applied the barycenter baseline on the French SemEval data (FSE) for verbs (Segonne et al., 2019) and on the FLUE benchmark for nouns (Le et al., 2020) to get a rough comparison (though datasets are quite different): for nouns, we reach comparable results (0.5353 accuracy), whereas the difference is quite large for verbs (0.5034 accuracy). One may partly explain this by the way annotated verbs were selected: medium frequency and medium rate of polysemy.

5.2 Analysis by degree of polysemy

Figure 1 shows the performance comparison for the different models in our experimental setup for disambiguating polysemic lemmas with respect to the number of senses per lemma. We observe that our proposed models tend to more effectively disambiguate polysemic lemmas with more than three-four senses than the MLP baseline (with some exceptions), showing the interest of using lexical network knowledge for those cases. For instance, for the verb *aller* (to go), our models predicted 8 distinct senses out of the 13 expected, while MLP baseline predicted 4 senses only.

6 Conclusion

We presented a preliminary study of various word sense disambiguation systems on the French dataset, DBLE-LN-fr-V2. We proposed a weighted training model in order to incorporate the richness of lexical and semantic information from the fr-LN network effectively and showed comparable performance to state of the art systems.

A first path of future research would be to enhance the scarcity of A matrix: e.g. adding neighbors of various POS, or including transitive clo-

sures of relations. We would like to explore the incorporation of sense embeddings using various graph representation learning algorithms. Furthermore, we would like to experiment tagset compression like in (Vial et al., 2019).

Acknowledgement

This work has been partly funded by the CPER Langues, Connaissances & Humanités Numériques (2014 - 2020).

References

Eneko Agirre, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art WSD. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 585–593, Sydney, Australia. Association for Computational Linguistics.

ATILF and CLLE. 2020. Corpus journalistique issu de l'est républicain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Ahmed El Sheikh, Michele Bevilacqua, and Roberto Navigli. 2021. Integrating personalized PageRank into neural word sense disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9092–9098, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Igor Mel'čuk. 2006. Explanatory Combinatorial Dictionary. In *Open Problems in Linguistics and Lexicography*, polimetrica, monza edition, pages 225–355.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

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

Alain Polguère. 2007. Lexical function standardness. In *Selected Lexical and Grammatical Issues in the Meaning–Text Theory*. John Benjamins.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Vincent Segonne, Marie Candito, and Benoît Crabbé. 2019. Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 259–270, Gothenburg, Sweden. Association for Computational Linguistics.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wroclaw, Poland. Global Wordnet Association.

Yorick Wilks and Mark Stevenson. 1998. Word sense disambiguation using optimised combinations of knowledge sources. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1398–1402, Montreal, Quebec, Canada. Association for Computational Linguistics.

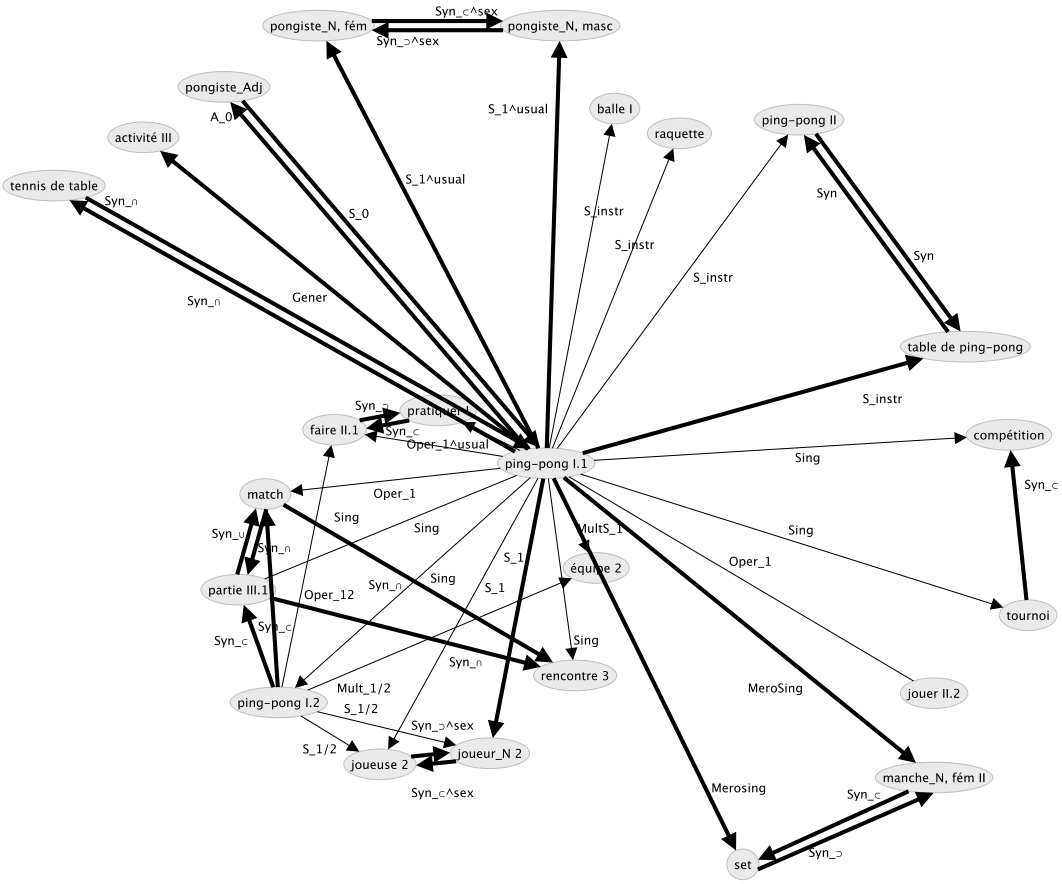

Figure 2: Extract of the *fr-LN* subgraph around the sense *ping-pong#I.1*. Only Lexical Function (LF) links are provided. The thickness of the lines reflects the semantic weight of the relation between two senses.

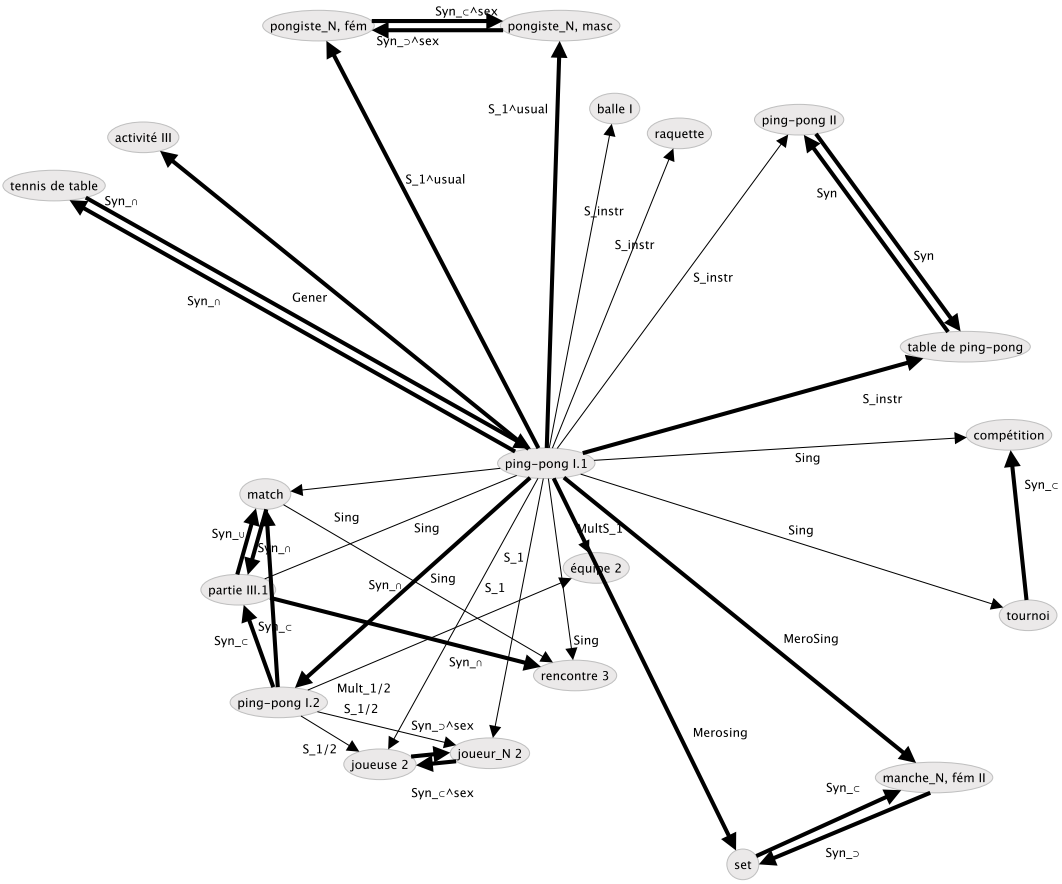

Figure 3: Extract of the *fr-LN* subgraph around the sense *ping-pong#I.1*. Only nouns and Lexical Function (LF) links are provided. The thickness of the lines reflects the semantic weight of the relation between two senses.