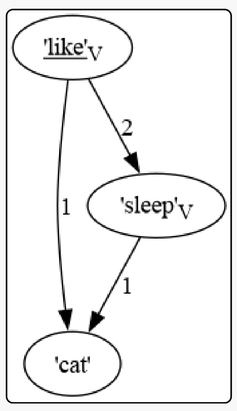
Flexible Lexicalization in Text Realization

Avril Gazeau, François Lareau
OLST, Université de Montréal

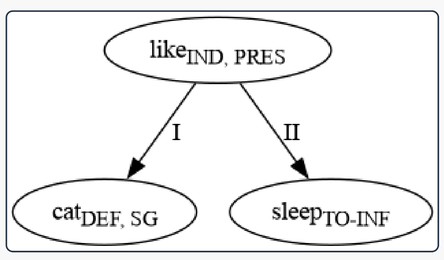
Université de Montréal

GenDR (Lareau et. al, 2018)

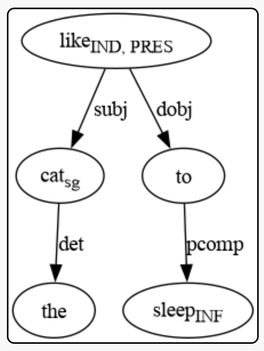
A symbolic deep text realizer based on Meaning-Text theory



GenDR takes a semantic structure (SemS) as input.



And transduces it into a deep syntactic structure (DSyntS), through a rule module and a semantic dictionary.



Then outputs a surface syntactic structure (SSyntS), through a different rule module and a lexical dictionary.

Deep lexicalization
Choosing the right lexical units (LUs) to express the SemS in a given language.

GenDR uses a **semantic dictionary (SD)** that maps semantic units to sets of corresponding lexicalizations.

Problem

- GenDR's current French SD contains only approx. 1,425 semantic units mapped to approx. 1,555 lexical units.
- It was compiled manually and contains errors and inconsistencies.

Objectives

- Create a module to automatically broaden the SD based on the French Lexical Network (LN-fr).
- Implement BERT in the module to further broaden the SD.
- Add a parameter to set the semantic distance between the SD entries and their lexicalizations, bringing more flexibility to GenDR.

The French Lexical Network (Polguère 2009, 2014)

The LN-fr is made up of nodes (LUs) and edges (lexical functions, or LFs).

LFs encode paradigmatic or syntagmatic relations. Semantically empty paradigmatic lexical functions (SEPLFs) encode substitution relations, thus their output has either exactly or approximately the same meaning as their input.

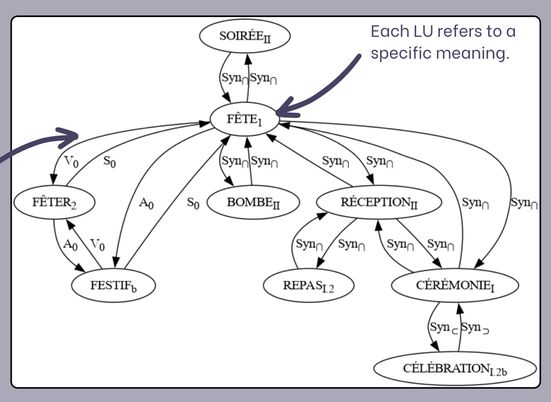


Figure: Sample of the LN-fr

Glosses

soirée$_{ii}$	'soirée'
fête$_{i}$	'party$_n$'
fêter$_{2}$	'party$_v$'
festif	'festive'
bombe$_{ii}$	'bash$_n$'
réception$_{ii}$	'function$_n$'
repas$_{i_2}$	'meal'
cérémonie$_{i}$	'ceremony'
célébration	'celebration'

LN-fr methodology

- Extract every node label from LN-fr to be an SD entry.
- Extract SEPLFs through RegEx.
- Annotate them as to whether they encode exact substitution (0) or approximate substitution (1).
- Add nodes linked to the entry through an SEPLF to the lexicalization set for said entry.
- Add lexicalizations recursively according to the Maximal Approximation Parameter (MAP).

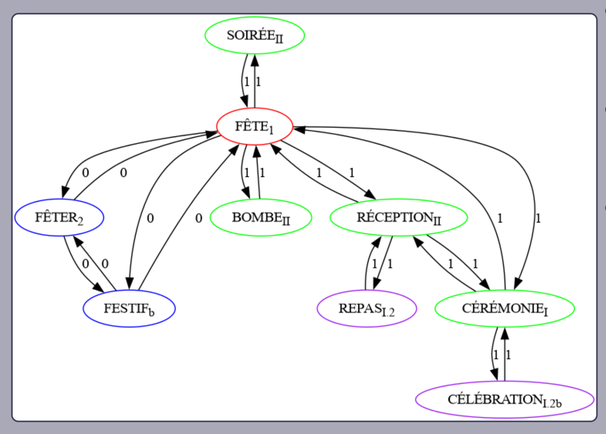


Figure: nodes added to the lexicalizations set for the entry 'fête$_1$' if MAP = 0, 1 or 2

- MAP is an integer.
- A counter is initialized to 0.
- Script explores nodes linked to the entry recursively.
- Approximation value of SEPLFs explored is added to counter.
- For every edge crossed, MAP and sum of counter are compared : if sum ≤ MAP, node is added as lexicalization. If sum > MAP, script goes back to entry, counter is reinitialized and script explores a new path.

Results

29 399 entries, min. 49 235 lexicalization links (SD 0), max. 572 686 lexicalization links (SD 5)

SD 0
party$_n$: {lex=party$_n$, lex=party$_v$, lex=festive}

SD 1
party$_n$: {lex=party$_n$, lex=party$_v$, lex=festive, qlex=bash$_n$, qlex=function$_n$, qlex=ceremony}

SD 2
party$_n$: {lex=party$_n$, lex=party$_v$, lex=festive, qlex=bash$_n$, qlex=function$_n$, qlex=ceremony, qlex=meal, qlex=celebration}

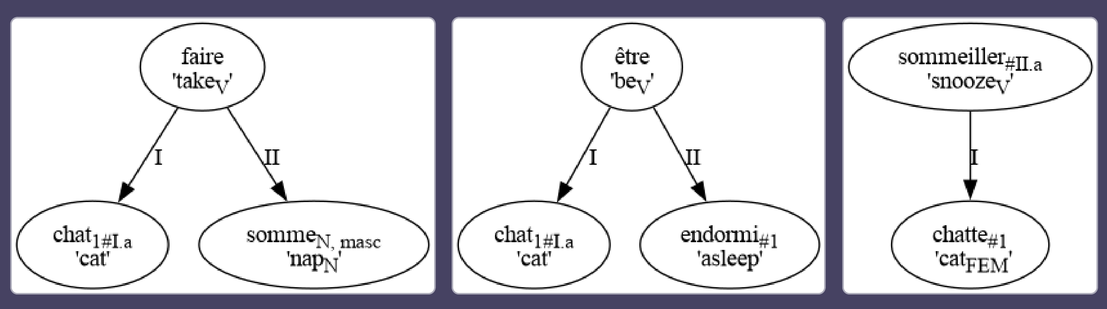


Figure: Small sample of DSyntSs generated using the SD 1 for the SemS 'The cat is sleeping'

CamemBERT (Martin et. al, 2020)

CamemBERT is a BERT model trained on French. BERT is trained on a Masked Language Modeling (MLM) objective and a Next Sentence Prediction (NSP) task using a special <SEP> token to separate sentences. LN-fr nodes are associated to 1+ example sentences and their offset is indicated.

Node	Sentence	Offset
manger$_{I.1a}$	Je [...] **mangeai** un sandwich de pain de mie à la tomate et au thon.	(42, 49)
'eat'	'I [...] **ate** a white-bread tomato and tuna sandwich.'	

We mask this offset and ask CamemBERT to return the most likely tokens. According to the distribution hypothesis, the tokens suggested by CamemBERT should have a similar meaning, thus are good candidates for lexicalizations of the SD entry with the same label as the node.

BERT methodology

- Method 1:

  I found an open bar, <MASK> a white-bread tomato and tuna sandwich.

- Method 2:

  I found an open bar, ate a white-bread tomato and tuna sandwich. <SEP> I found an open bar, <MASK> a white-bread tomato and tuna sandwich.

10 candidates with their certainty scores between 0 and 1 were generated per node for both methods.

Results

Entry	Lexicalizations	Candidates
country	{kingdom, place}	{homeland, paradise, cradle, land, kingdom, 'this one', capital}
armoured	{reinforced}	{medieval, military, dating, french, gothic, defensive, royal, urban, german, strong}
unforgettable	{immortal, memorable}	{relentless, memorable, epic, terrible, intense, bloody, singular, hard, final}

Table 1: Sample of candidates generated with CamemBERT

Evaluation

- Comparing candidate sets with lexicalization sets for same SD entry. If CamemBERT can reproduce the data compiled by lexicographers, chances are the candidates are good enough to enhance the SD.
- Precision, recall and F-score according to different normalized certainty score thresholds to see if higher certainty means higher quality.
- According to SDs produced with MAP ranging 0-5 to see if semantic distance has impact on CamemBERT's capacity to reproduce SD data.

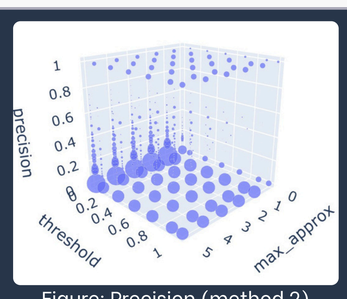


Figure: Precision (method 2)

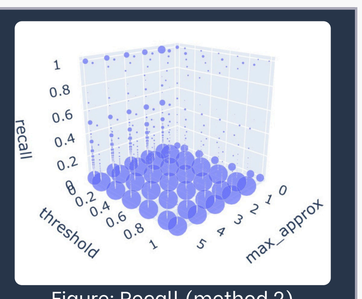


Figure: Recall (method 2)

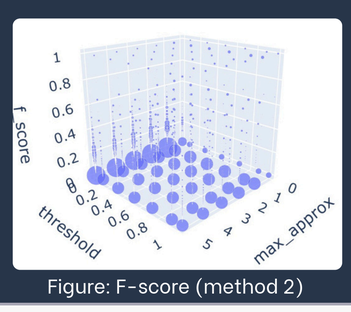


Figure: F-score (method 2)

Evaluation (continued)

- Manual evaluation of sample of candidates absent from SD 0 and 1 to see if they can truly enhance it.
- Candidates with normalized certainty scores ≥ 0.85 only
- 500 candidates per SD (~1% of total generated candidates per SD.

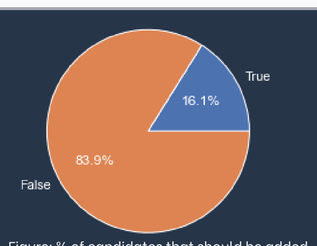


Figure: % of candidates that should be added to SD 0 (method 2)

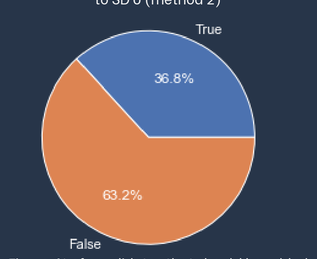


Figure: % of candidates that should be added to SD 1 (method 2)

Conclusion

- We successfully created a flexible lexicalization module that can produce SDs with various degrees of semantic accuracy compatible with GenDR. In doing so, we identified a new type of LF, the SEPLFs, that had not yet been mentioned in the literature.
- We were unable to use CamemBERT to enhance the SD because of its low precision, recall and F-score against the SD. Although the manual evaluation shows more promising results, especially with the second method, the amount of acceptable candidates produced is still too low to introduce them systematically in the SD without adding a lot of noise as well.

References

- Bohnet, B., Langjahr, A. & Wanner, L. (2000). A development environment for an MTT-based sentence generator. In *Proceedings of the first international conference on Natural language generation (INLG)*, p. 260, Mitzpe Ramon, Israel.
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota.
- Lareau, F., Lambrey, F., Dubinskaitė, I., Galarreta-Piquette, D. & Nejat, M. (2018). GenDR: A generic deep realizer with complex lexicalization. In *Proceedings of 11th edition of the Language Resources and Evaluation Conference (LREC)*, Miyazaki.
- Mel'čuk, I. & Polguère, A. (2021). Les fonctions lexicales dernier cri. In Marengo, S. (ed.), *La Théorie Sens-Texte. Concepts-clés et applications*, pp. 75–155. L'Harmattan.
- Mel'čuk, I. (1981). Meaning-Text Models: A Recent Trend in Soviet Linguistics. *Annual Review of Anthropology*, 10(1):27–62.
- Polguère, A. (2009). Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*, 43(1):41–55.
- Polguère, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4):396–418.