Expériences de désambiguïsation lexicale

Soutenance de stage - Clémentine BLEUZE

L3 MIASHS Sciences Cognitives, année universitaire 2021-2022

Sommaire

1) L’ATILF
2) Matériel informatique, logiciels et langages
3) Sujet du stage et contexte
4) Travail réalisé
5) Conclusion

1. L’ATILF

- **A**nalyse et **T**raitement **I**nformatique de la **L**angue **F**rançaise
- linguistique et informatique

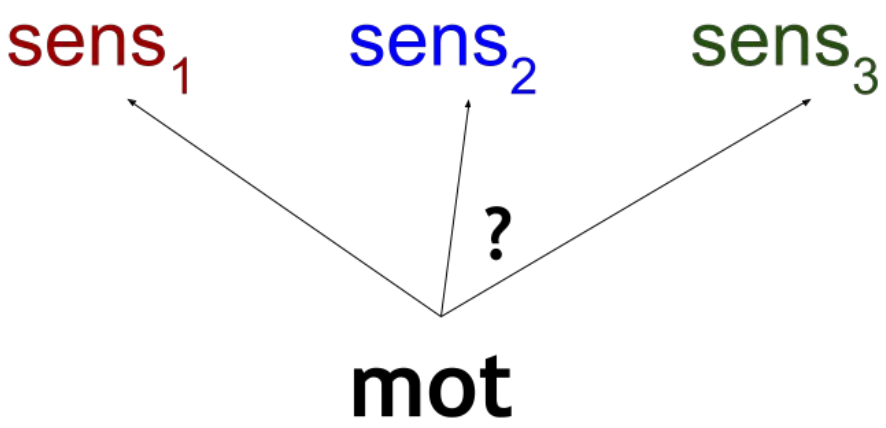
- CNRS + UL

- étude du lexique, aspects dynamiques du langage, corpus et ressources

2. Matériel informatique, logiciels et langages

3. Sujet du stage et contexte

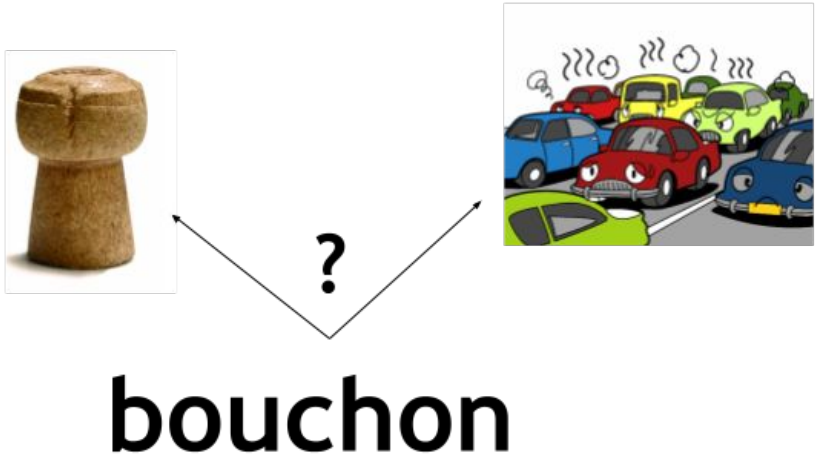
> La désambiguïsation lexicale



- tâche de classification d'une occurrence de mot parmi ses sens possibles

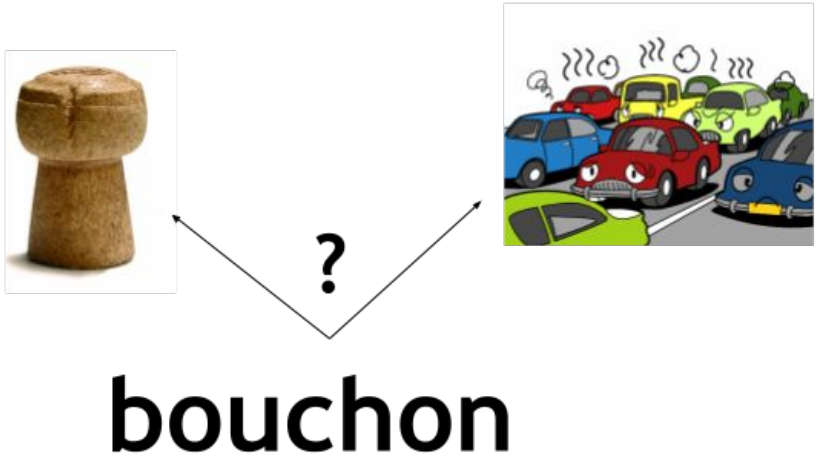
3. Sujet du stage et contexte

> La désambiguïsation lexicale



3. Sujet du stage et contexte

> La désambiguïsation lexicale



“ Il est coincé dans les **bouchons**.”

3. Sujet du stage et contexte

> La désambiguïsation lexicale

?

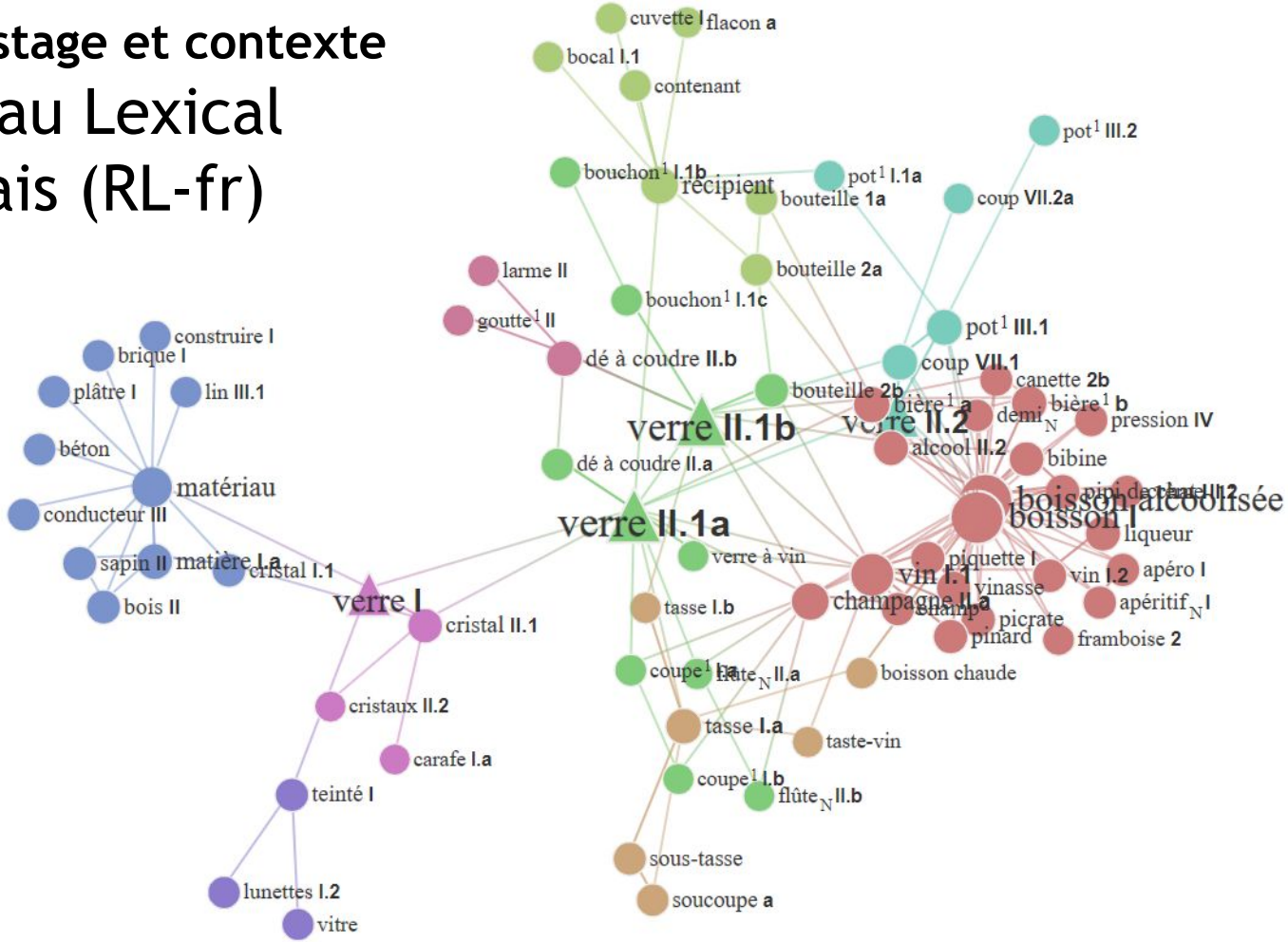
bouchon

“ Il est coincé dans les **bouchons**.”

- on doit disposer au préalable d’un **inventaire de sens** pour “bouchon”

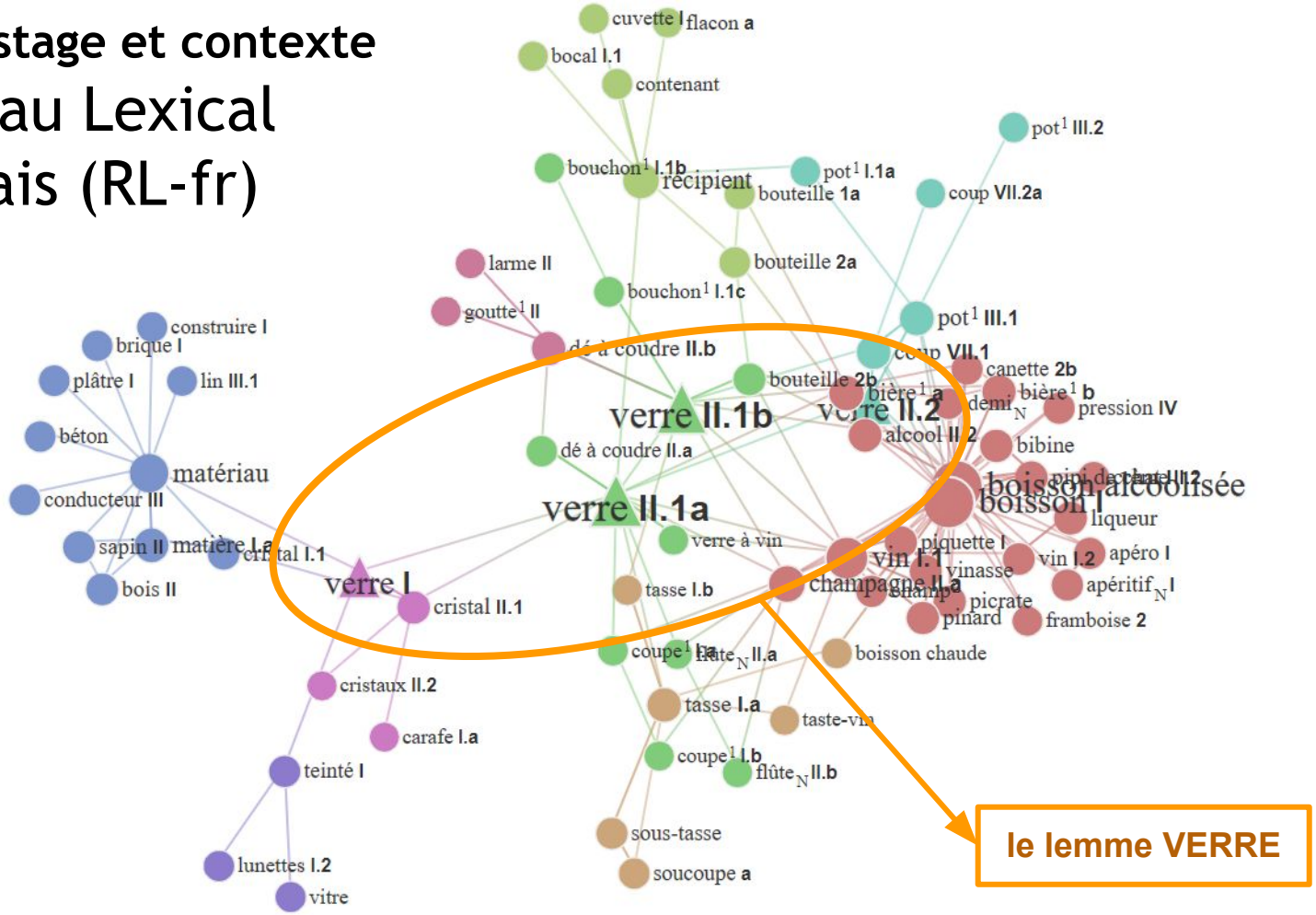
3. Sujet du stage et contexte

> Le Réseau Lexical du français (RL-fr)



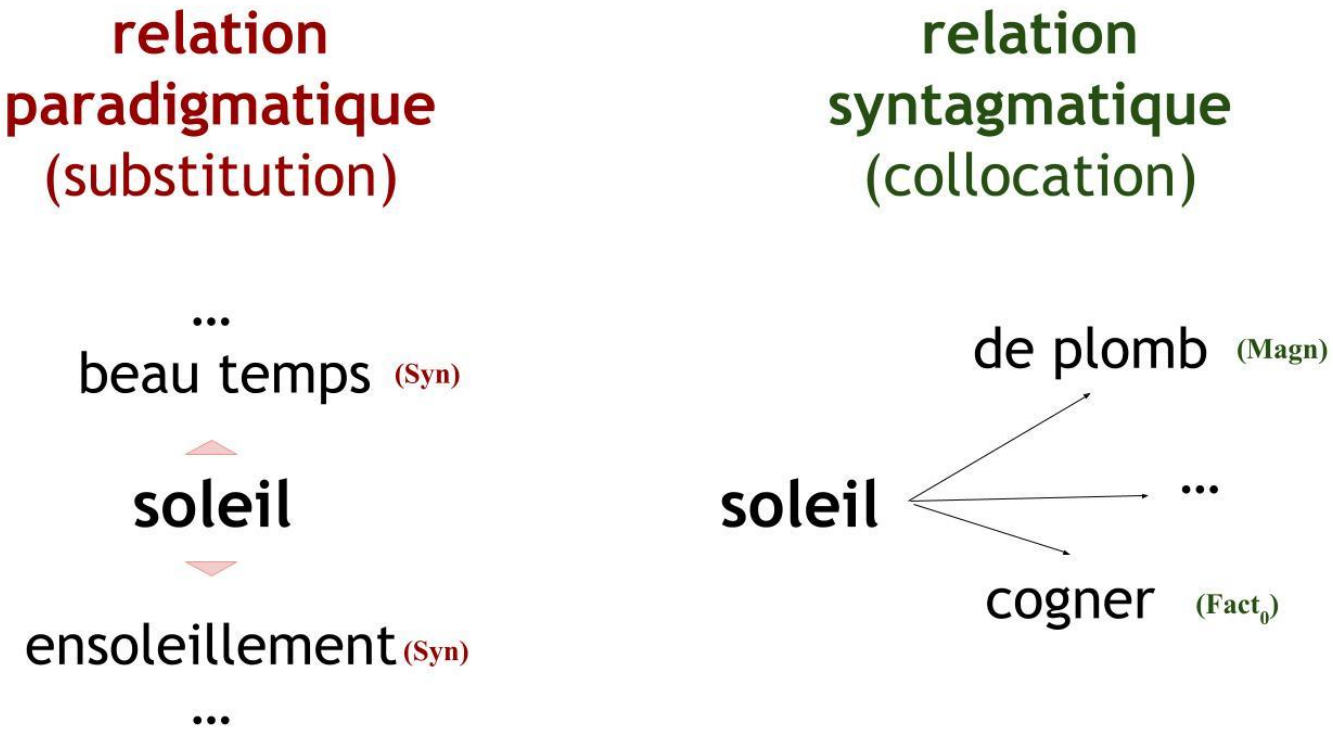
3. Sujet du stage et contexte

> Le Réseau Lexical du français (RL-fr)



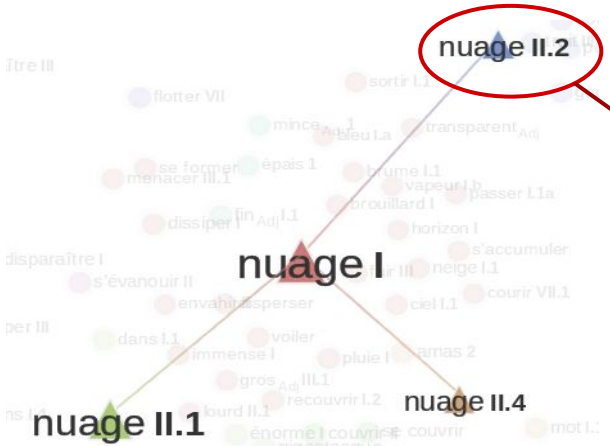
3. Sujet du stage et contexte

> Le RL-fr



3. Sujet du stage et contexte

> Le BEL-RL-fr



[EX]

Une goutte de miel, un **nuage** de lait, une pinte de sang dans la mer Méditerranée ne sauraient troubler sa couleur uniformément vineuse.

Frantext SERRES Michel, *Le Tiers-Instruit*, 1991, p. 232

Au moment du rituel du thé versé dans les tasses, **nuages** de lait, un sucre ?

Frantext NAVARRE Yves, *Biographie*, 1981, p. 684

Je bois deux tasses de thé avec un **nuage** de lait, repousse ma chaise perpendiculairement et pondérément par rapport à la table (mais sans atteindre le dos de la chaise située derrière moi) et murmure perspicacement à l'intention du couple japonais : « Rain today, I'm afraid ! »

Frantext ROUBAUD Jacques, *Le Grand Incendie de Londres : récit, avec incises et bifurcations*, 1989, p. 231

Malgré le lobe, papa nous a autorisés à inaugurer le premier des trente pots de crème de marrons qu'il vient de faire, dont quinze sont additionnés d'un **nuage** de crème au chocolat.

Frantext GROULT Benoîte, GROULT Flora, *Journal à quatre mains*, 1994, p. 107

Ce sont les Anglais qui, à partir du début du 19ème siècle, ont donné aux indiens l'habitude de mettre du thé noir et le fameux « **nuage** de lait ».

FrWac 02 2008, http://www.yogitea.fr/yogitea/spip.php?rubrique32

3. Sujet du stage et contexte

> Le BEL-RL-fr

[EX]

Une goutte de miel, un **nuage** de lait, une pinte de sang dans la mer Méditerranée ne sauraient troubler sa couleur uniformément vineuse.

Frantext SERRES Michel, *Le Tiers-Instruit*, 1991, p. 232

Au moment du rituel du thé versé dans les tasses, **nuages** de lait, un sucre ?

Frantext NAVARRE Yves, *Biographie*, 1981, p. 684

Je bois deux tasses de thé avec un **nuage** de lait, repousse ma chaise perpendiculairement et pondérément par rapport à la table (mais sans atteindre le dos de la chaise située derrière moi) et murmure perspicacement à l'intention du couple japonais : « Rain today, I'm afraid ! »

Frantext ROUBAUD Jacques, *Le Grand Incendie de Londres : récit, avec incises et bifurcations*, 1989, p. 231

Malgré le lobe, papa nous a autorisés à inaugurer le premier des trente pots de crème de marrons qu'il vient de faire, dont quinze sont additionnés d'un **nuage** de crème au chocolat.

Frantext GROULT Benoîte, GROULT Flora, *Journal à quatre mains*, 1994, p. 107

Ce sont les Anglais qui, à partir du début du 19ème siècle, ont donné aux indiens l'habitude de mettre du thé noir et le fameux « **nuage** de lait ».

FrWac 02 2008, http://www.yogitea.fr/yogitea/spip.php?rubrique32

- distribution ORTOLANG
- citations illustrant 1 à 12 lexies
- v2 (2021): +30 000 citations et +50 000 associations citation-lexie

→ corpus de citations lexicographiques partiellement désambiguïsé

3. Sujet du stage et contexte

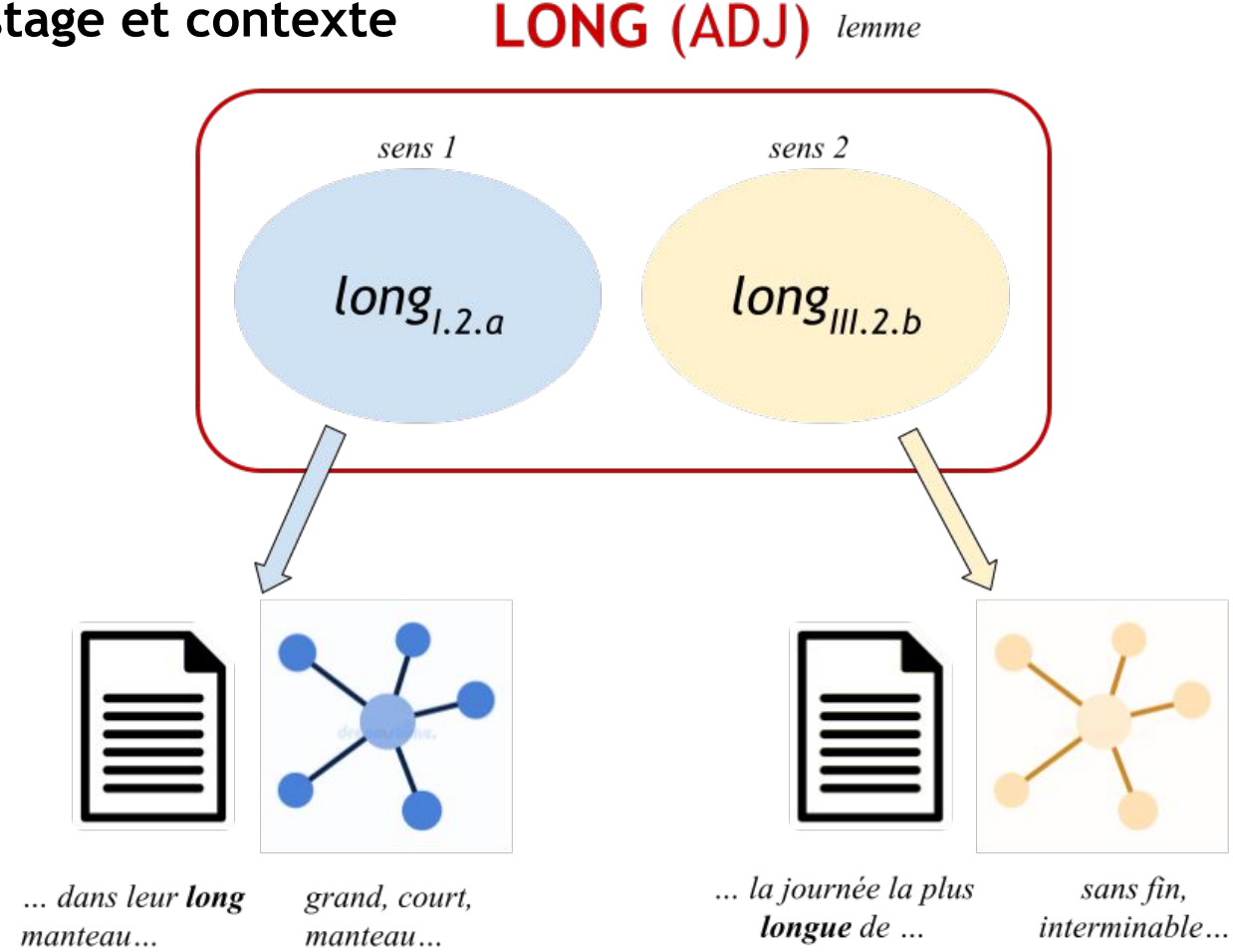
- 25 **lemmes** (50 **sens)** issus de l'inventaire du RL-fr

Objectif = mettre en place plusieurs méthodes de désambiguïsation basiques pour ces couples

Pour chaque sens:

- des phrases exemples
- les voisins lexicaux (niveau 1) du RL-fr

3. Sujet du stage et contexte



3. Sujet du stage et contexte

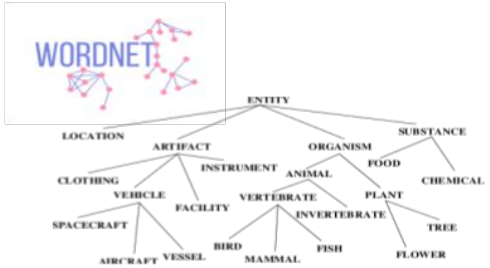
> 25 couples à désambiguïser

- **adjectifs** : long
- **prépositions** : avec
- **adverbes** : ailleurs, près
- **noms** : antenne, boucher, bouchon, éclair, fauteuil, lessive, lit, marche, rouge, soleil, veste
- **verbes** : aller, allonger, chanter, courir, dormir, manger, marcher, nettoyer, sortir, voler

3. Sujet du stage et contexte

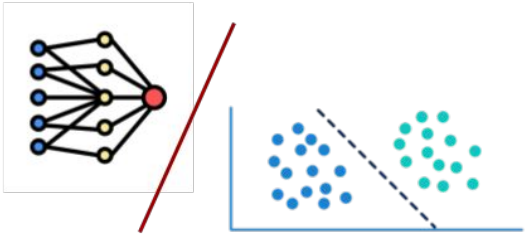
> Enjeux et applications

RL-fr vs. autres ressources lexicales



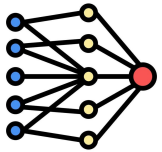
Enrichissement et annotation du BEL-RL-fr

Confrontation à des expériences avec réseaux de neurones



3. Sujet du stage et contexte

> Expériences de désambiguïsation

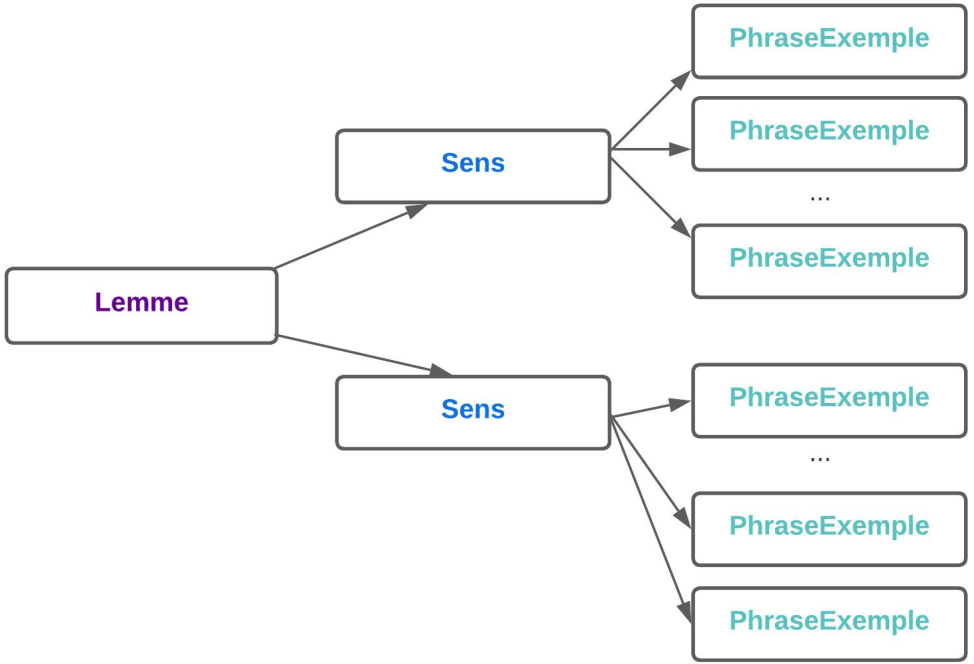
Réseaux de Neurones (Constant, Ollinger, Sinha)	**Méthodes de désambiguïsation basiques, classifieurs**
	
+27 000 énoncés	+3 000 énoncés
BEL-RL-fr	BEL-RL-fr + frTenTen17
classification parmi n sens (n_{max} = 32, pour les verbes)	classification parmi 2 sens

4. Travail effectué

> Modélisation et pré-traitements

- Programmation Orientée Objet (Python)

- pré-traitements avec **UDPipe**



4. Travail effectué

> Pré-traitements sur les données

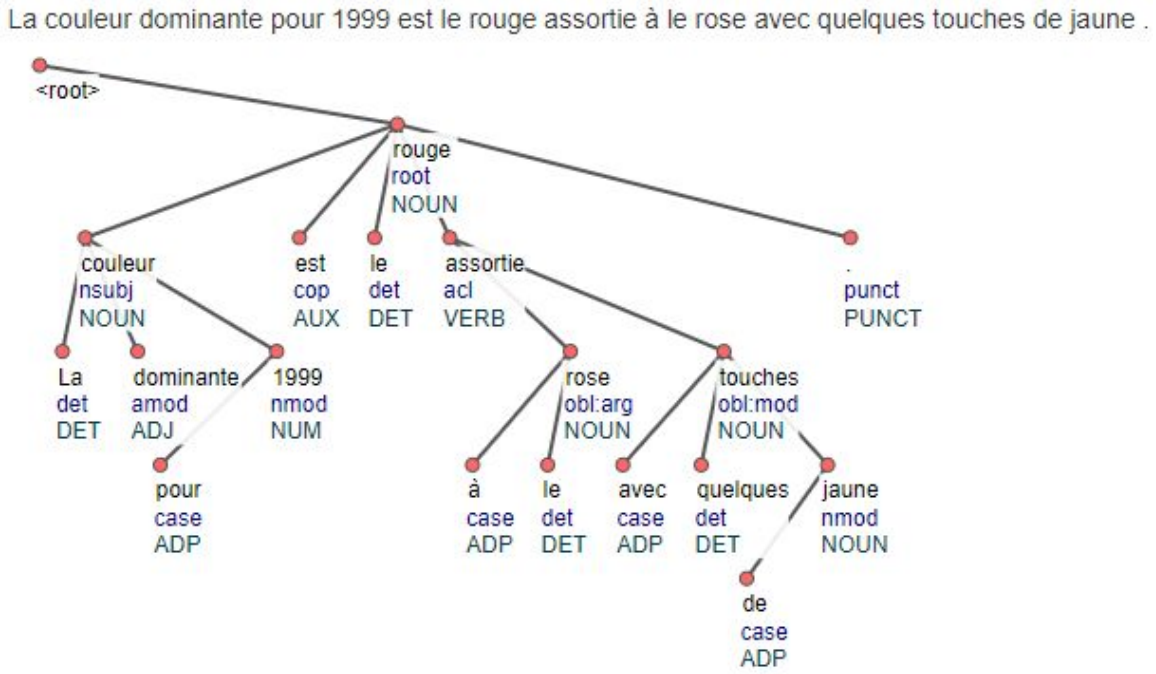
- **analyse syntaxique UDPipe:** tokénisation, lemmatisation, POS-tagging …

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# newdoc									
# newpar									
# sent_id = 1									
# text = La couleur dominante pour 1999 est le rouge assortie au rose avec quelques touches de jaune.									
1	La	le	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	2	det	_	TokenRange=0:2
2	couleur	couleur	NOUN	_	Gender=Fem|Number=Sing	8	nsubj	_	TokenRange=3:10
3	dominante	dominant	ADJ	_	Gender=Fem|Number=Sing	2	amod	_	TokenRange=11:20
4	pour	pour	ADP	_	_	5	case	_	TokenRange=21:25
5	1999	1999	NUM	_	_	2	nmod	_	TokenRange=26:30
6	est	être	AUX	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	8	cop	_	TokenRange=31:34
7	le	le	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	8	det	_	TokenRange=35:37
8	rouge	rouge	NOUN	_	Gender=Masc|Number=Sing	0	root	_	TokenRange=38:43
9	assortie	assortier	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	8	acl	_	TokenRange=44:52

4. Travail effectué

> Pré-traitements sur les données

- **analyse syntaxique UDPipe:** arbre de dépendances syntaxiques



4. Travail effectué

> Méthodes de désambiguïsation

- méthode aléatoire — *baseline*

- méthode des voisins
- méthode des cooccurrents

— *conçues à la main*

- classifieur Naive Bayes
- classifieur Régression Logistique
- classifieur Random Forest

— *classifieurs de référence en TAL*

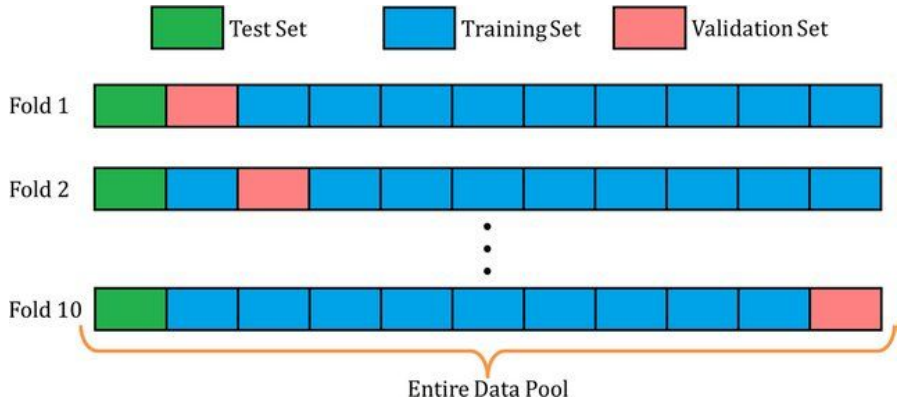
4. Travail effectué

> Méthodologie

1) Mise au point de la méthode / documentation

2) Implémentation en Python

3) Test (phase de développement) et premiers résultats



précision = 100 * nombre de prédictions correctes / nombre de prédictions effectuées

4. Travail effectué

> Méthodologie

1) Mise au point de la méthode / documentation

2) Implémentation en Python

3) Test (phase de développements) et premiers résultats

4) Observation des résultats et comparaison avec les autres méthodes

4. Travail effectué

> Exemple: méthode des voisins

1) Mise au point de la méthode / documentation

ROUGE

```
Voisins du sens 'I.1b': ['couleur', 'sombre', 'foncé', 'vif', 'brillant', 'ardent', 'pâle', 'clair', 'vermillon', 'rubis', 'pourpre', 'nacarat', 'lie-de-vin', 'incarnat', 'grenat', 'écarlate', 'carmin', 'bordeaux', 'virer', 'passer']

Voisins du sens 'IV': ['qui tache', 'gros', 'vin rouge', 'vin', 'blanc', 'boisson', 'boisson alcoolisée']
```

4. Travail effectué

> Exemple: méthode des voisins

1) Mise au point de la méthode / documentation

ROUGE

```
Voisins du sens 'I.1b': ['couleur', 'sombre', 'foncé', 'vif', 'brillant', 'ardent', 'pâle', 'clair', 'vermillon', 'rubis', 'pourpre', 'nacarat', 'lie-de-vin', 'incarnat', 'grenat', 'écarlate', 'carmin', 'bordeaux', 'virer', 'passer']

Voisins du sens 'IV': ['qui tache', 'gros', 'vin rouge', 'vin', 'blanc', 'boisson', 'boisson alcoolisée']
```

“La couleur dominante pour 1999 est le rouge assorti au rose, avec quelques touches de jaune.”

→ sens : $\text{rouge}_{\text{I.1.b}}$

4. Travail effectué

> Exemple: méthode des voisins

1) Mise au point de la méthode / documentation

ROUGE

```
Voisins du sens 'I.1b': ['couleur', 'sombre', 'foncé', 'vif', 'brillant', 'ardent', 'pâle', 'clair', 'vermillon', 'rubis', 'pourpre', 'nacarat', 'lie-de-vin', 'incarnat', 'grenat', 'écarlate', 'carmin', 'bordeaux', 'virer', 'passer']

Voisins du sens 'IV': ['qui tache', 'gros', 'vin rouge', 'vin', 'blanc', 'boisson', 'boisson alcoolisée']
```

“La couleur dominante pour 1999 est le rouge assorti au rose, avec quelques touches de jaune.”

→ sens : $\text{rouge}_{\text{I.1.b}}$

“Ce vin possède une belle robe d'un rouge sombre.”

→ sens: ???

4. Travail effectué

> Exemple: méthode des voisins

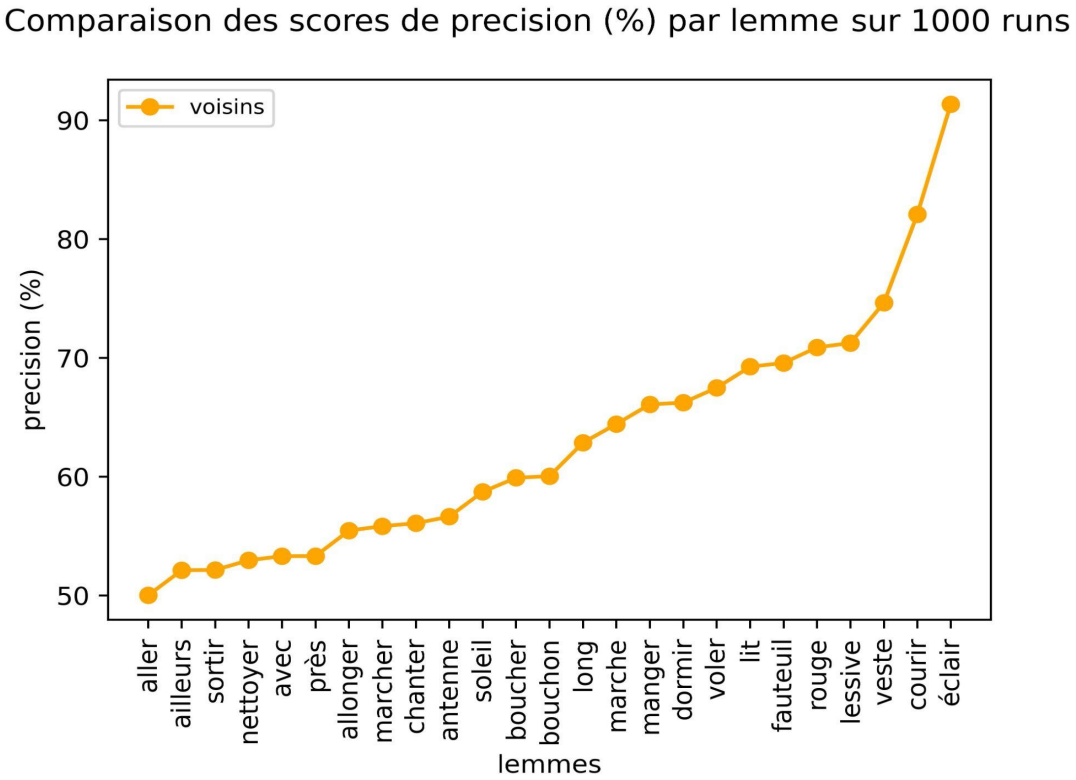
2) Implémentation en Python

4. Travail effectué

> Exemple: méthode des voisins

3) Test (phase de développements) et premiers résultats

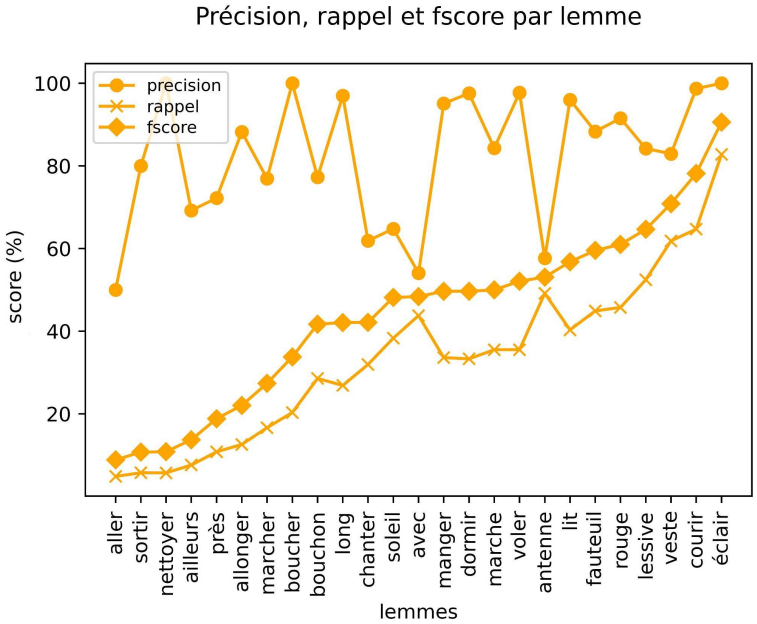
- précision moyenne: 63%



4. Travail effectué

> Exemple: méthode des voisins

4) Observation des résultats et comparaison avec les autres méthodes



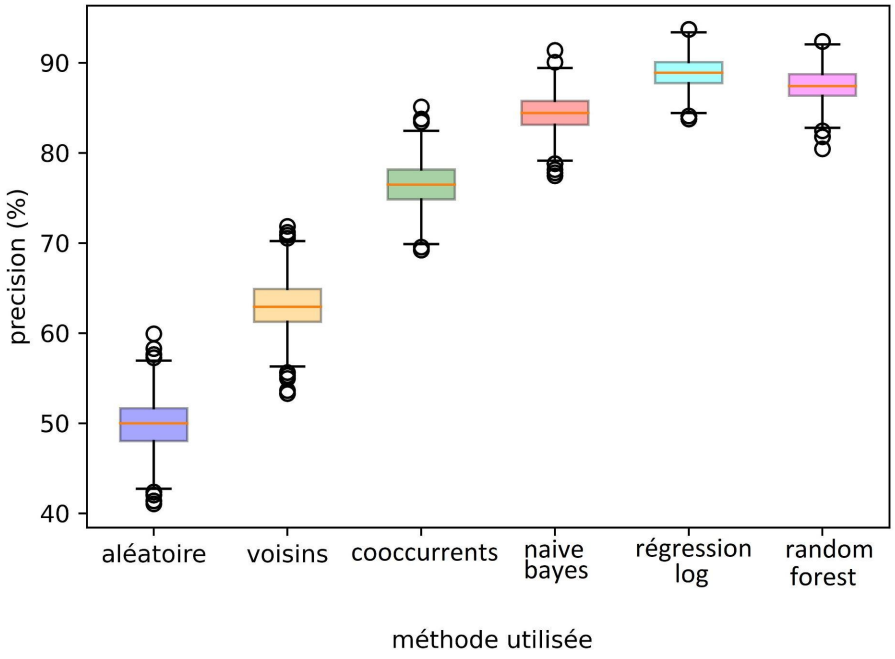
- version sans aléatoire

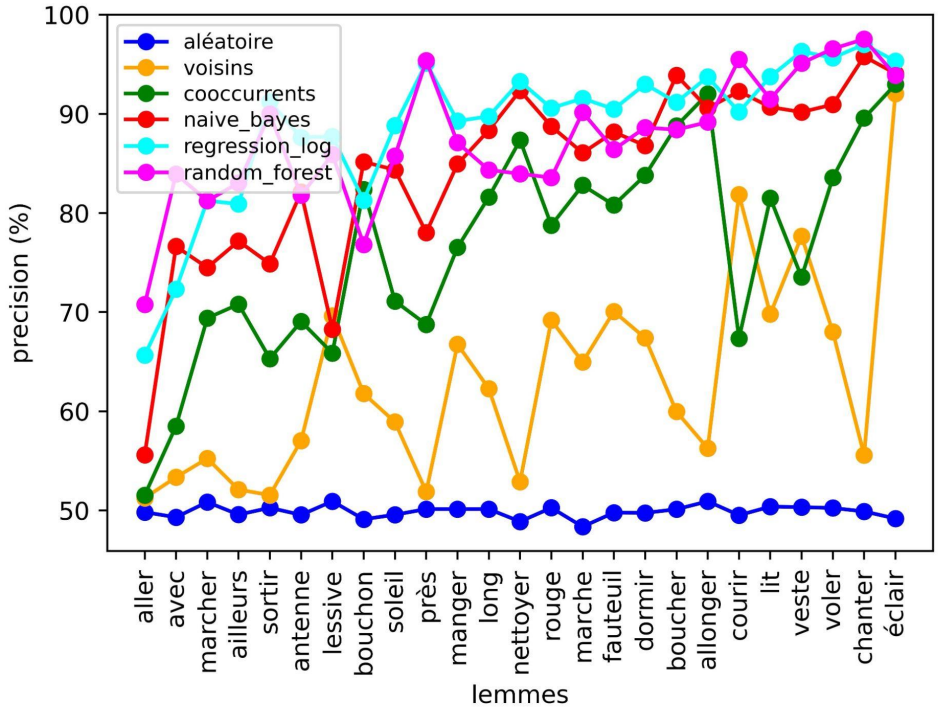
- seulement 41% des énoncés prédits
- égalités: quasiment toujours à 0 voisins

- précision = 81.54%
- rappel = 33.26%
- f-score = 47.25 % < 50%

4. Travail effectué

> Comparaison des méthodes





4. Travail effectué

> Choix d'une représentation de données

Les modèles ne travaillent pas directement sur les données d'entrée, mais sur une représentation de ces données.

Un énoncé ~ un vecteur

V = vocabulaire appris
F = features (caractéristiques de la représentation)

v_i	arbre	avoir	bien	ce	…	nuit	pouvoir	sable	zèbre
f_i	0	1	1	1	…	1	0	0	0

Cette nuit a été bien longue. → **(0 1 1 1 … 1 0 0 0)**

4. Travail effectué

> Choix d’une représentation de données

Conception d’une représentation de données pour la Régression Logistique:

- proposition de critères
  → vocabulaire, catégories grammaticales, entités nommées, voisins lexicaux…

- proposition de contextes
  → global
  → local/ locaux

- comparaison progressive critère par critère, contexte par contexte

4. Travail effectué

> Choix d’une représentation de données

	nb de run	précision moyenne	variance	durée (min)
Expé 1	1000	80.37%	5.04	~30
Expé 2	1000	81.01%	4.25	~50
Expé 3	1000	84.03%	4.05	~30
Expé 2a	100	81.73%	5.24	~5
Expé 2b	100	79.87%	5.24	~5
Expé 2c	100	78.70%	3.15	~5
Expé 2d	100	81.04%	4.12	~5

4. Travail effectué

> Choix d'une représentation de données

critère	procédé d'évaluation du critère	C_t	C_g	C_d
voisins RL-fr	décompte des voisins de chaque sens dans C			
mots du vocabulaire	fréquence d'apparition dans C			
voisins syntaxiques	décompte des voisins de chaque sens dans C			
entités nommées	décompte des EN de chaque catégorie dans C			
catégories grammaticales	décompte des CG par catégorie dans C			

4. Travail effectué

> Choix d'une représentation de données

	nb_voisins_0_g	nb_voisins_0_d	nb_voisins_0_t	nb_voisins_1_g	nb_voisins_1_d	nb_voisins_1_t	nb_voisins_synt_0_g	nb_voisins_synt_0_d
0	0.0	0.333333	0.250000	0.0	0.0	0.000000	0.333333	0.333333
1	0.0	0.000000	0.055556	0.0	0.0	0.055556	0.666667	0.666667
2	0.0	0.000000	0.000000	0.0	0.0	0.047619	0.333333	0.000000
3	0.0	0.000000	0.111111	0.0	0.0	0.000000	0.500000	0.666667
4	0.0	0.000000	0.062500	0.0	0.0	0.000000	0.666667	0.000000

nb_voisins_synt_0_t	nb_voisins_synt_1_g	...	acceptable_g	acceptable_d	acceptable_t	2008_g	2008_d	2008_t	plaisir_g	plaisir_d	plaisir_t	id_sens
0.500000	0.333333	...	0	0	0	0	0	0	0	0	0	0
0.555556	0.666667	...	0	0	0	0	0	0	0	0	0	0
0.142857	0.666667	...	0	0	0	0	0	0	0	0	0	1
0.555556	0.500000	...	0	0	0	0	0	0	0	0	0	0
0.562500	0.333333	...	0	0	0	0	0	0	0	0	0	0

4. Travail effectué

> Scores finaux

méthode	paramètres	toutes catégories (25)	noms (11)	Verbes (10)
aléatoire	/	**49,74**	49,53	49,64
voisins	avec aléatoire	61,25	**64,98**	57,58
voisins	sans aléatoire	44,71	**54,5**	28,47
cooccurrents	avec aléatoire, nb_cooc = 25	76,56	76,32	**79,58**
cooccurrents	sans aléatoire, nb_cooc = 25	72,08	72,02	**73,31**
naive_bayes	/	83,44	84,09	**86,07**
regression log	Set 230	88,74	87,12	**92,5**
random forest	Set 230	87,09	86,52	**88,5**

4. Travail effectué

> Analyse qualitative

lemme	**ECLAIR**	**ALLER**	**SORTIR**	**SOLEIL**
caractéristiques	meilleurs résultats	pires résultats	facile pour l'humain mais difficile pour les modèles	déjà étudié
relation de copolysémie	homonymie	métaphore	métaphore	métonymie

4. Travail effectué

> Analyse qualitative

lemme	**ECLAIR**	**ALLER**	**SORTIR**	**SOLEIL**
caractéristiques	meilleurs résultats	pires résultats	facile pour l'humain mais difficile pour les modèles	déjà étudié
relation de copolysémie	homonymie	métaphore	métaphore	métonymie

- sélection de 50 énoncés

4. Travail effectué

> Analyse qualitative

données	- relecture des énoncés

4. Travail effectué

> Analyse qualitative

données	- relecture des énoncés
pré-traitements	- anlyse UDPipe: vérification * tokénisation * lemmatisation * catégorie grammaticale - vérification de la bonne identification des occurrences

4. Travail effectué

> Analyse qualitative

données	- relecture des énoncés
pré-traitements	- anlyse UDPipe: vérification * tokénisation * lemmatisation * catégorie grammaticale - vérification de la bonne identification des occurrences
application des méthodes	- voisins et cooccurrents: retrouver les résultats, repérer les indices non-captés ou les faux indices captés - observation des features importants - lien avec la relation de copolysémie

4. Travail effectué

> Analyse qualitative

données	- relecture des énoncés
pré-traitements	- anlyse UDPipe: vérification * tokénisation * lemmatisation * catégorie grammaticale - vérification de la bonne identification des occurrences
application des méthodes	- voisins et cooccurrents: retrouver les résultats, repérer les indices non-captés ou les faux indices captés - observation des features importants - lien avec la relation de copolysémie
amélioration des méthodes	- tentative d'amélioration de la méthode des voisins, en intégrant le voisinage lexical de niveau 2

4. Travail effectué

> Analyse qualitative

- vérification analyse UDPipe: tokénisation, lemmatisation, POS-tagging

type de défaut	éclair	aller	sortir	soleil
tokénisation	5	2	4	0
lemmatisation	3	0	17	1
POS	4	2	13	4
total	12	4	34	5

→ comparer avec d'autres outils ?
→ source d'erreur variable selon les lemmes

4. Travail effectué

> Analyse qualitative

- observation manuelle des prédictions par voisins, cooccurrents

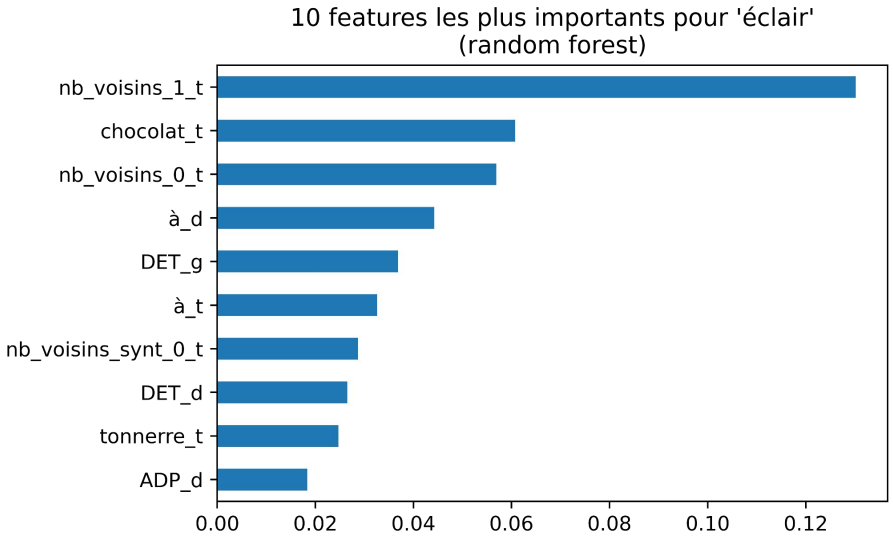
→ peu de prédictions

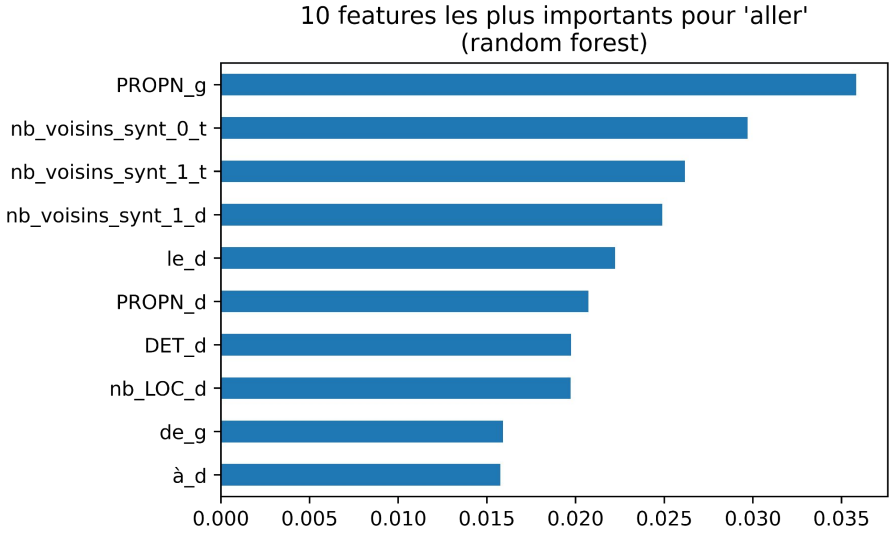
→ des voisins ou cooccurrents captés par erreur (“se rendre compte”)
→ des indices manqués

4. Travail effectué

> Analyse qualitative

- features importants (régression logistique et random forest)





4. Travail effectué

> Analyse qualitative

Méthode des voisins

- Constats:
  - trop grande part de l’aléatoire
  - les voisins n’apparaissent pas assez en contexte
  - disproportions entre les voisins des différents sens

4. Travail effectué

> Analyse qualitative

Méthode des voisins

- Constats:
  - trop grande part de l'aléatoire
  - les voisins n'apparaissent pas assez en contexte
  - disproportions entre les voisins des différents sens

- Idée:
  - enrichir la liste des voisins grâce aux voisins lexicaux de **niveau 2**
  - prendre en compte la nature de la relation avec les voisins (paradigmatique et syntagmatique)

4. Travail effectué

> Analyse qualitative

nb de prédictions correctes / effectuées	éclair	aller	sortir	soleil
niveau 1 seulement	7/7	1/1	1/1	4/5
niveau 2 syntagmatiques seulement	5/5	5/8	6/12	1/1
niveau 2 paradigmatiques seulement	10/10	1/2	2/2	7/12
niveau 2 syntagmatiques & paradigmatiques	**11/11**	5/9	6/12	7/12

- amélioration des résultats pour **“éclair”**
- pour les autres: on augmente le nombre de prédictions mais la précision chute.

5. Conclusion

Merci pour votre attention