

BEL-RL-fr

Sandrine Ollinger

Expériences de désambiguïsation lexicale

STRange Thursdays

27 janvier 2022

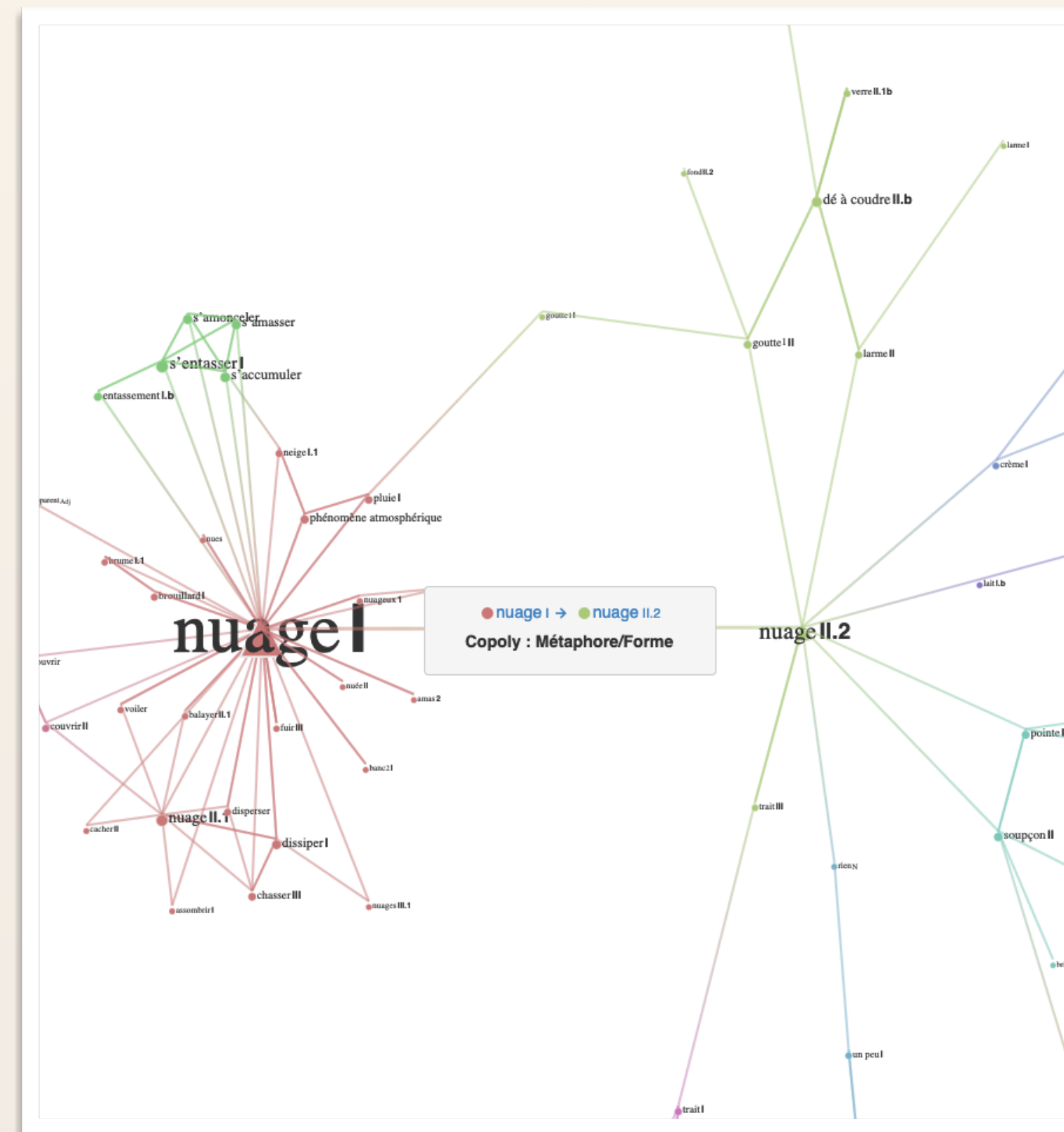
CONTEXTE

RL-fr

- Travail lexicographique depuis 2011
- 1 unité lexicale =
1 forme + 1 sens + 1 combinatoire
- Travail sur la *polysémie*
- Représentation **fine** du sens
- 29 334 unités lexicales
- 18 664 vocables
- 1,57 unités lexicales par vocable
- 80 611 liens entre unités lexicales

TLFi

- 100 000 mots (5x plus)
- 270 000 définitions (9x plus)



EXEMPLES

- 1 exemple = 1 citation + 1 segment textuel associé au sens illustré
- Chaque unité lexicale illustrée à l'aide de 1 à n exemples (max 58)
- DB : table citations + table associations
- 33 874 citations
- 52 699 associations
- 1 association peut contenir plusieurs segments textuels

TLFi

- 430 000 exemples (8x plus)

[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 19^{ème} 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>

id	source_id	status	text	title	authors	location	date_day	date_
8791	1	1	<html><body> Une goutte de miel, un nuage de lait, une pinte de sang dans la mer Méditerranée ne sauraient troubler sa couleur uniformément vineuse. </body></html>	Le Tiers-Instruit	Serres, Michel	p. 232	NULL	

id	sense_id	ex_id	position	enabled	percent	comments	occurrences
10658	42532	8791	0	1	100	<html><body> </body></html>	38,68;

CITATIONS

CHANGEMENT DE POINT DE VUE

- 1 citation = court texte associé à 1 à n unités lexicales (max 12)
- BEL-RL-fr = corpus de citations partiellement désambiguïsées en unités lexicales du RL-fr
- 33 874 citations
- 28 147 unités lexicales
- 54 739 occurrences

id	sense_id	ex_id	position	enabled	percent	comments	occurrences
10658	42532	8791	0	1	100	<html><body> </body></html>	38,68;
17854	44400	8791	0	1	100	<html><body> </body></html>	19,50;
27475	37207	8791	0	1	100	<html><body> </body></html>	129,166;

ID : 8791	
Classe : Citations de corpus type Frantext	
Source : Frantext	
Statut : En attente de validation (1)	
Citation	
Une goutte de miel, un nuage de lait, une pinte de sang dans la mer Méditerranée ne sauraient troubler sa couleur uniformément vineuse.	
Référence	
Frantext SERRES Michel, <i>Le Tiers-Instruit</i> , 1991, p. 232	
Occurrences	
goutte ¹ II	Une goutte de miel, un nuage de lait, une pinte de sang dans la mer Méditerranée ne sauraient troubler sa couleur uniformément vineuse.
nuage II.2	Une goutte de miel, un nuage de lait, une pinte de sang dans la mer Méditerranée ne sauraient troubler sa couleur uniformément vineuse.
uniformément	Une goutte de miel, un nuage de lait, une pinte de sang dans la mer Méditerranée ne sauraient troubler sa couleur uniformément vineuse.
Création par seha le 6 mars 2013 à 19:18:55	
Modification par seha le 6 mars 2013 à 19:18:55	

Désambiguïser davantage

MAIS POUR QUOI FAIRE ?

INTÉRÊT APPLICATIONS

TAL & PÉDAGOGIE

- RL-fr exploité en TAL et dans applications pédagogiques
- Les exemples présents :
 - Pas toujours en nombre suffisant
 - Pas toujours adaptés à la tâche
- Souhait de pouvoir propager la désambiguïsation à d'autres citations :
 - Dans la base elle-même
 - */soleils?/* → 244 occurrences dont seulement 45 dans segments
 - Dans d'autres corpus, pour fournir d'autres exemples
 - Pour la description lexicographique
 - Pour les applications pédagogiques
 - Pour le TAL

INTÉRÊT RL-FR

& SYSTÈMES LEXICAUX

- Descriptions sont-elles correctes ?
- La polysémie a-t-elle bien été dégagée ?
- Y a-t-il des liens de copolysémie « dynamiques » ?

Je regarde les **corbeaux** en me disant qu'ils sont beaux et qu'ils brillent au soleil.

(https://forum.doctissimo.fr/psychologie/reves-cauchemars/idee-interpretation-cauchemard-sujet_184788_1.htm, 24/01/22)

- CORBEAU#I.a [Le **corbeau** (*Corvus corax*) est un grand oiseau noir de la famille des corvidés.]
- CORBEAU#I.b [Les gens parlent de **corbeaux**, mais il peut s'agir de freux, de corneilles.]
 - ▶ Pas besoin de trancher entre CORBEAU#I.a et CORBEAU#I.b
- CORBEAU#II.2b [C'est un **corbeau**. Elle écrit des lettres anonymes.]
 - ▶ Mais il faut exclure CORBEAU#II.2b

INTÉRÊT AMBIGUÏTÉ

ÉTUDE DE L'AMBIGUÏTÉ LEXICALE

- Comprendre le phénomène
- Cas simples vs. cas complexes vs. cas à ne pas désambiguïser
- Rapports entre ambiguïté et polysémie
- L'ambiguïté comme ressort langagier



PARTICULARITÉS

TÂCHE DE DÉSAMBIGUISATION AUTOMATIQUE

- Approche semi-automatique
- Ressource lexicale incomplète (en cours de constitution)
- Faible quantité de données
- Genre textuel (citations lexicographiques)
- Désambiguïser une forme + flexions et non toutes les formes d'un texte (*Lexical sample* [Navigli 2009])

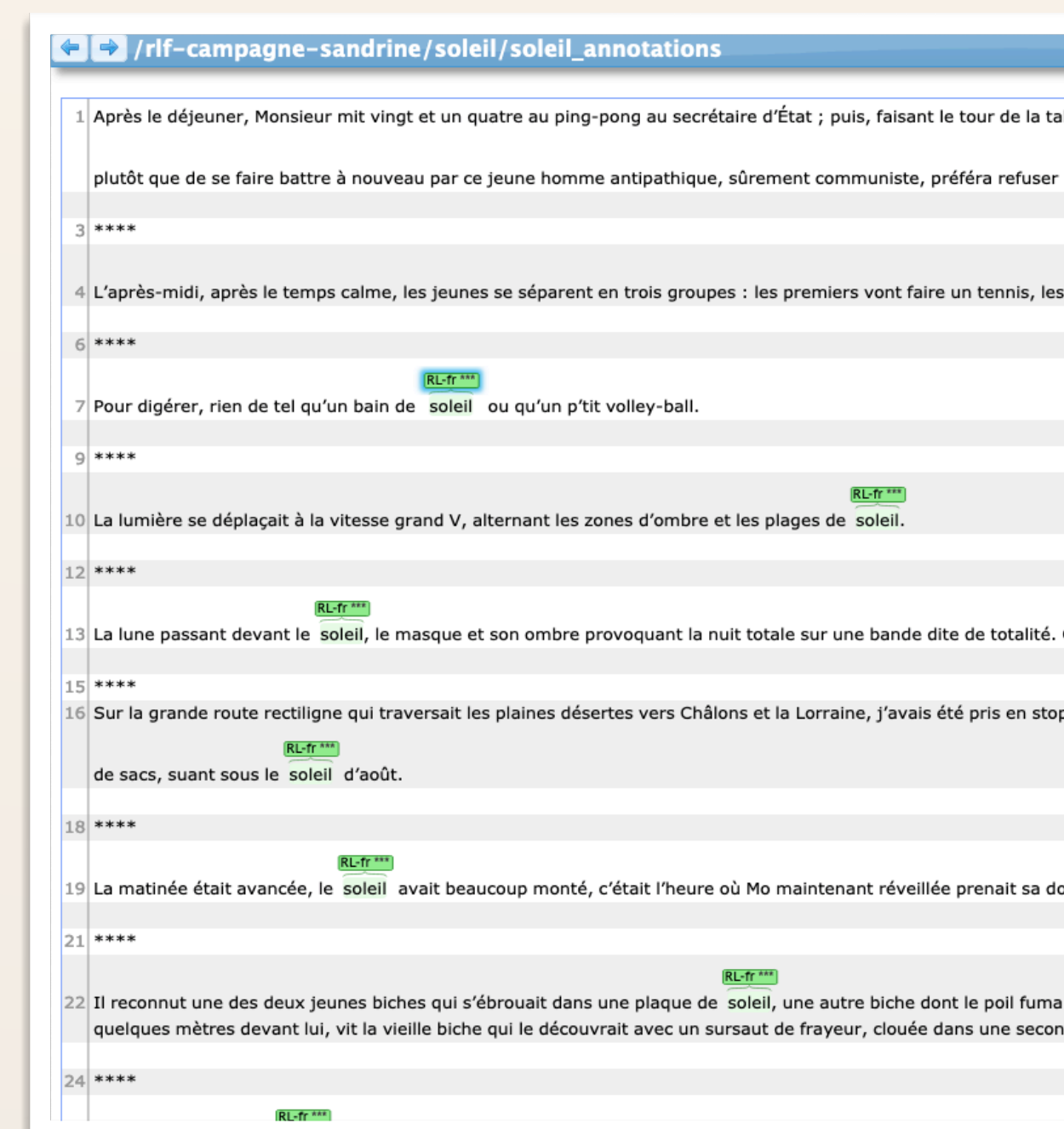
TRAVAUX PASSÉS

BEL-RL-fr de 2018 à 2021

CoReA2D

MARS — OCTOBRE 2018

- Avec : Évelyne Jacquey, Veronika Lux, Simon Méoni, Alain Ploguère, Marie-Sophie Pausé
- Financement : CPER LCHN
- Annotation manuelle (BRAT) de segments de citations en unités lexicales :
 - 4 unités lexicales AUBE
45 citations
 - 6 unités lexicales SOLEIL
244 citations
- 4 annotateurs
- 2 arbitres



Export XML

SEPTEMBRE — DÉCEMBRE 2019

- Avec : Marine Schmitt, Étienne Petitjean (coup de pouce Cyril Pestel, Jessika Perignon)
- Financement : CORLI + CPER LCHN
- Export BEL-RL-fr format XML TEI
- Chaîne de traitement reproductible (GitLab)
- Distribution Ortolang V1

```
<cit xml:id="cit4124" type="example" source="ls:fr:ex:4175" ana="#FrW">
  <quote>L'aigle <seg xml:id="seg10229" type="st000000000000" ana="ls:fr:gc:23"
function="active" source="https://spiderlex.atilf.fr/fr/id/29871">vole</seg> seul ; ce
sont les <seg xml:id="seg10230" type="st000000000000" ana="ls:fr:gc:20" function="active"
source="https://spiderlex.atilf.fr/fr/id/26912">corbeaux</seg>, les choucas et les
étourneaux qui <seg xml:id="seg10231" type="st000000000000" ana="ls:fr:gc:23"
function="active" source="https://spiderlex.atilf.fr/fr/id/26268">vont</seg> en
groupe...</quote>
  <bibl>
    <title/>
    <author/>
    <date when="2008-02">02 2008</date>
    <ref
type="url">http://kohlanta.tf1.fr/kohlanta/forum/0,,3478042-
e01FU19JRCBVTlZfSUR9IHsXNjQ20TMwIDF9,00-forum-koh-lanta-.html</ref>
  </bibl>
</cit>
```

- * FÉVRIER 2020 : Export TXM
- * AVRIL 2021 : Distribution Ortolang V2

Validation

JUIN — AOÛT 2020

- Avec : Camille Kunz, Alain Polguère
- Financement : IMPACT OLKi-LUE
- Validation manuelle des citations (17%)
 - Typographie
 - Orthographe
 - Longueur
 - Clarté
 - Biblio
 - Marques d'usage ?
- Report dans la base en cours
- Rédaction mémo Système Lexicaux
Exemples lexicographiques

```
=====
ls:fr:ex:306
Dame·A·Diomandé·résident·au·KM·17·(dans·les·encablures

Titre :
Auteur :
Date : 07·01·2013
Emplacement : http://news.abidjan.net/h/447393.html
=====

Il faudrait peut être plus anonymiser les noms propres
il faut une espace insécable entre 17 et heures
il faut une espace insécable entre 20 et millions
il faut une espace insécable entre KM et 17 ??
exemple peut être un peu long
=====
ls:fr:ex:313
Son·dossier·serait·sans·doute·resté·enfoui·dans·les·arc

Titre :
Auteur :
Date : 02·10·2012
Emplacement : http://www.bladi.net/forum/25135-patrouil
=====
```



TRAVAUX EN COURS

2020 — 202?

Phrases

JANVIER 2020 — ??

- Avec : Denis Maurel (Lifat), Nathalie Friburger (Lifat)
- Cascade segmentation en phrases UNITEX
- V1 BEL-RL-fr
- Sous-corpus dev : 19 892 citations
- Sous-corpus test : 10 000 citations
- Cas retors : discours rapporté, dialogues, phrases imbriquées

```
<div xml:id="cit4131">
  <p><s>L'aigle vole seul ; ce sont les corbeaux, les choucas et les étourneaux qui
    vont en groupe...</s></p>
</div>
<div xml:id="cit4132">
  <p><s>Un escargot quelle drôle de p'tite bête.</s>
  <s>Alors bien sûr tu peux les ramasser dans le jardin.</s>
  <s>Mais tu peux aussi en fabriquer des tout doux.</s>
  <s>Il te faut juste un cure-pipe, des perles suffisamment larges pour les
    enfiler sur le cure-pipe.</s></p>
</div>
<div xml:id="cit4133">
  <p><s>Des mouettes et des canards volaient ça et là et des cormorans tout noirs
    plongeaient en quête de nourriture tandis que quelques cygnes se pavanaient
    sur les rives du fleuve.</s></p>
</div>
<div xml:id="cit4134">
  <p><s>Son corps aérodynamique était joliment dessiné, avec une petite tête, les
    ailes et la queue longues et pointues.</s>
  <s>Cette silhouette permettait à l'oiseau de voler vite, avec grâce et
    agilité.</s></p>
</div>
<div xml:id="cit4135">
  <p><s>Cette mouche a une flottaison très basse et l'utilisation du dubbing de lièvre
    à poils longs donne la vie à ce montage.</s></p>
</div>
<div xml:id="cit4136">
  <p><s>Il faut nettoyer sa pipe après chaque fumage, aussi bien intérieurement
    qu'extérieurement.</s>
  <s>Le nettoyage du tuyau se fait avec des nettoie-pipes, aussi appelés cure-pipe
    (ce sont des brins de métal recouverts de fibres absorbantes).</s></p>
</div>
<div xml:id="cit4137">
  <p><s>Ils traversaient une clairière quand Samuel s'exclama :</s>- <s>Nom d'une pipe
    !</s>
  <s>J'ai un pneu à plat !</s>- <s>J'ai de quoi réparer dans la sacoche.</s></p>
</div>
```

Bert & Co

NOVEMBRE 2020 — ??

- Avec : Aman Sinha, Mathieu Constant, Alain Polguère
- Expériences de désambiguisation lexicale automatique
- Approche supervisée
- Réseaux de neurones
- Plongements lexicaux (FlauBERT, CamemBERT)
- Verbes & Noms
- Segments annotés BEL-RL-fr V2 répartis en train-dev-test
- Golds + Inspiration EWISER (Navigli)
- Fichiers prédiction fraîchement arrivés

Model Description	VERB	NOUN
MFS-Baseline	0.14884	0.20837
LFS-Baseline	0.10796	0.20364
RS-Baseline w.o inventory	0.11740	0.19198
RS-Baseline w. inventory	0.16561	0.24444
KNN sent-level	0.00733	0.01238
KNN-all sent-level	0.03563	0.03788
KNN word-level	0.14465	0.29289
KNN-all word-level	0.27882	0.45646

- 9 050 occurrences de verbes
- 2 446 lemmes
- 5 141 sens
- 26 024 occurrences de noms
- 7 968 lemmes
- 13 059 sens

QUESTIONNEMENTS

DÉSAMBIGUISATION AUTOMATIQUE

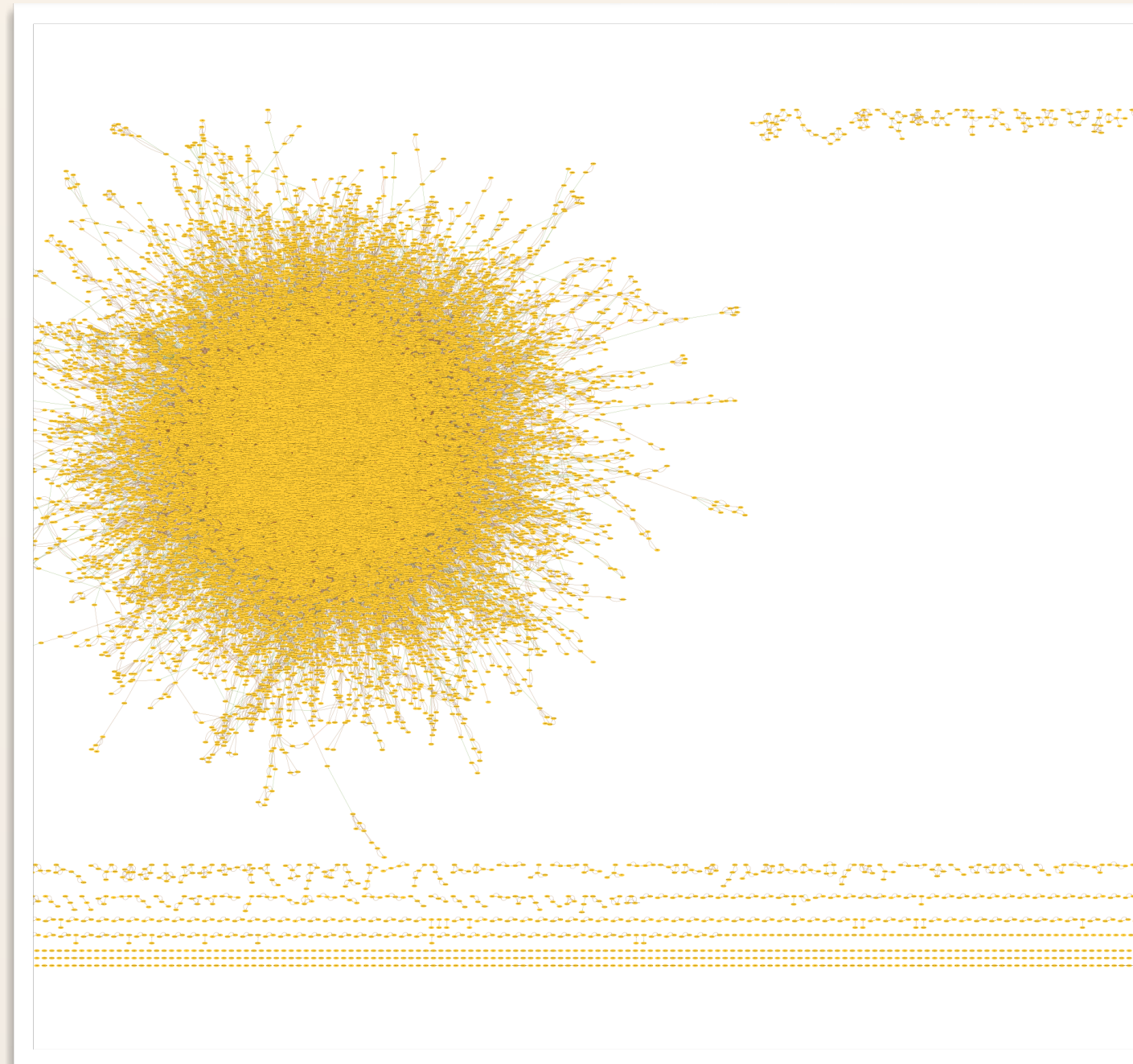
- Quelle alternative à la présence de définitions ?
- Quelle alternative à l'évaluation binaire des résultats ?
 - Si j'attends CORBEAU#I.a, CORBEAU#I.b est MOINS FAUX que CORBEAU#II.2b
- Mesure de distance dans le graphe lexicale ? (Confluence ?)

TRAVAUX À VENIR

2022 — 202?

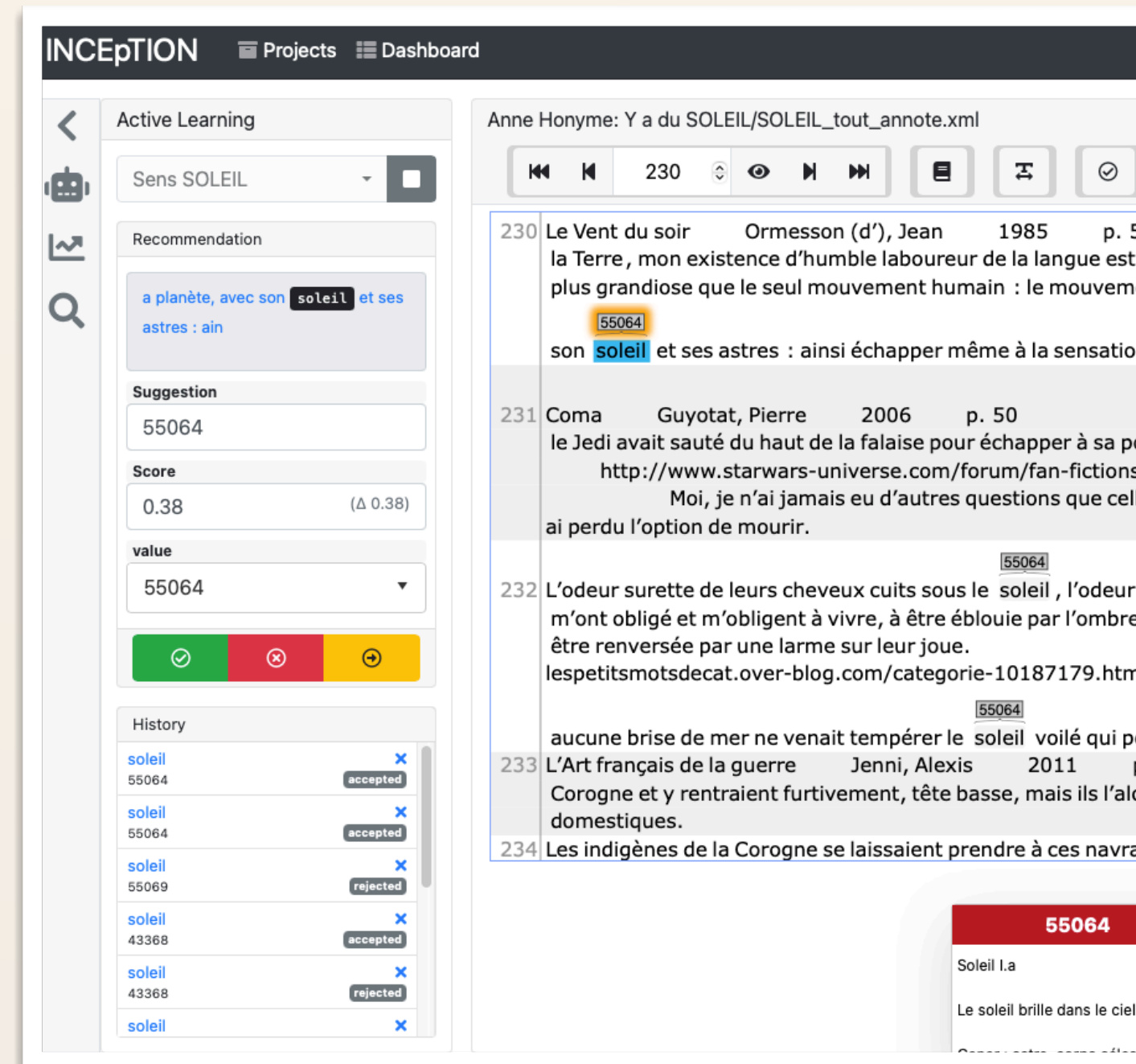
Graphes

- Avec : Bruno Gaume (CLE-ERSS), Pierre Magistry (INALCO), ...
- Évaluation non binaire
- Prise en compte de la structure du graphe dans apprentissage
- Groupe de travail en visio



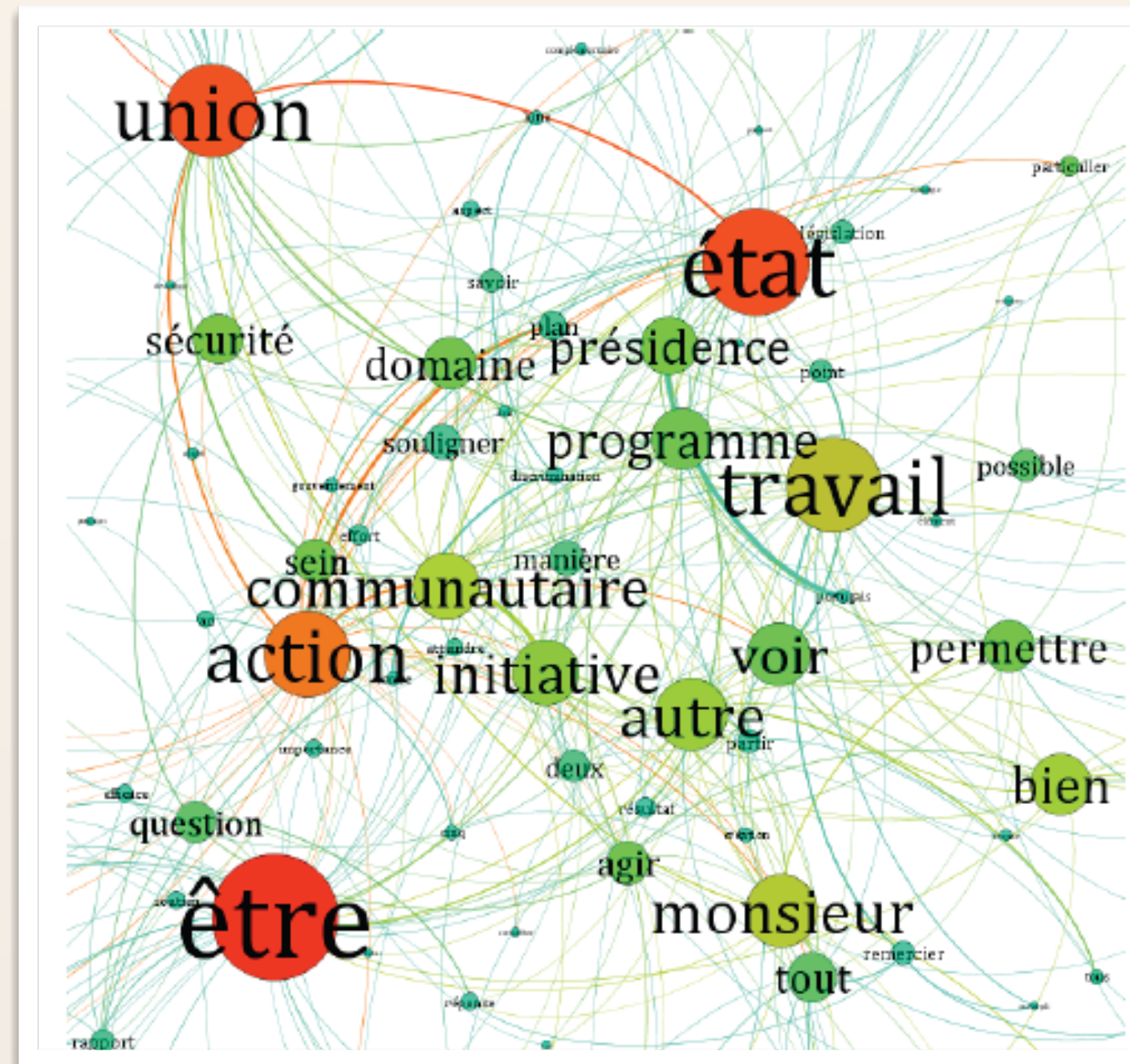
INCePTION

- Avec : Nabil Gader, Évelyne Jacquey, Veronika Lux, Alain Polguère, ...
- Annotation manuelle
- Logiciel INCePTION
- Couplé avec désambiguïsation automatique



Gold étendu

- Avec : Clémence Bleuze, Mathieu Constant
- Stage L3 MIASHS
- 23 mai – 13 août
- État de l'art



Tauveron 2012

GT interne

- Avec : Évelyne Jacquey, Alain Polguère, Veronika Lux, Mathieu Constant
- Groupe de travail *Ambiguïté lexicale*
- Projet laboratoire 2023-2027
- Compléter avec expériences sur TLFi





MERCI