



Outils pour la veille néologique

Grands types de néologismes

- ✓ Deux distinctions bien connues et transversales

	Formelle	Sémantique
Terminologie	X	X
Langue générale	X	X

- ✓ La distinction formelle / sémantique régit les classifications des néologismes
 - *Formels* : apparition de nouvelles formes par des mécanismes liés à l'évolution des formes des vocables (dérivation, composition, emprunt)
 - *Sémantiques* : apparition de nouveaux sens liés aux sens pré-existants de vocables existant déjà (extension/restriction de sens ; analogie/métaphore ; contiguïté/métonymie)
- ✓ Automatiser la détection des néologismes = veille néologique

Deux approches de la veille néologique

- ✓ Une approche traditionnelle s'appuyant sur les méthodologies de la lexicographie
 - *Relevés (veille néologique)*
 - *Validation, analyse des néologismes candidats et rédaction de fiches d'attestation*
- ✓ Une approche émergente s'appuyant sur les méthodologies de la linguistique textuelle et du discours
 - *Relevés (veille néologique)*
 - *Classification des néologismes candidats en fonction de leur(s) contexte(s) discursif(s) et thématique(s)*

Deux approches de la veille néologique

- ✓ **Des approches qui se complètent**
 - *L'approche lexicographique va jusqu'à la description complète des néologismes*
 - *L'approche textuelle et discursive systématise l'utilisation des classifications thématiques et discursives*
- ✓ **Des approches qui se rejoignent**
 - *L'approche lexicographique se dote d'outils (semi-)automatisés pour la validation des NC et la réalisation des fiches d'attestation*
 - *L'approche lexicographique s'inspire de l'approche textuelle et discursive pour l'identification et l'homogénéisation des thématiques des documents où apparaissent les néologismes*
- ✓ **Linguistique de corpus et classification thématique au service de la veille néologique pour les néologismes sémantiques**
 - *Forme existante apparaissant avec de nouvelles thématiques et/ou un nouveau profil distributionnel et combinatoire => Néologisme sémantique candidat*

Deux approches : deux outils de veille

✓ Néoveille (ancien NEOLOGIA) // LIPN : 2014

> ...

□ *Cartier, Sablayrolles, Humbley et al. 2016 et 2018, Cartier 2019*

□ *Veille néologique en flux continu sur 7 langues contemporaines (plus de 250 sources)*

✓ Logoscope // LILPA : 2014 - 2018

□ *Gérard, Falk, Bernhard 2014*

□ *Veille néologique en flux continu sur le français dans la presse quotidienne (10 journaux)*

Néoveille : accès

- ✓ Théoriquement accessible sous forme d'une plateforme web : <https://www.neoveille.org>
 - ❑ *Mais en fait non. Demande à E. Cartier restée sans réponse à ce jour*
- ✓ Accessible via github
<https://github.com/ecartierlipn/neoveille2016>
 - ❑ *Github un peu trop compliqué => à rediscuter avec l'équipe de développement au laboratoire*

Néoveille : structure



Gestion des sources



Détection NC



Évaluation NC



BD documentée de néologismes

Néoveille : gestion des sources

- ✓ Identification des sources à ajouter / modifier / supprimer (flux RSS / sites web)
- ✓ Encodage de métadonnées
 - *journal, URL, public visé, domaine, langue, pays*
- ✓ Récupération automatisée des articles 2X/j
 - *Extraction de métadonnées de documents*
 - × titre, auteur(s), date de publication, mots-clés, domaine
 - *Extraction du contenu textuel, étiquetage morphosyntaxique et segmentation en phrases*

Néoveille : vue synthétique des sources

- ✓ Distributions selon le temps, le pays, le domaine, les journaux



Néoveille : détection des néologismes candidats (NC)

✓ NC = Identification des mots inconnus et application de filtres

- ❑ *Dictionnaires de référence*
- ❑ *Dictionnaires d'exclusion*
- ❑ *Identification des noms propres*
- ❑ *Identification des erreurs typographiques*

Néoveille : module de validation des NC

✓ Interface graphique en ligne

- ❑ Liste des candidats à vérifier individuellement par chaque expert sous forme de tableau
- ❑ Validation ou non, type de néologismes (typologie Sablayrolles 2016), commentaires

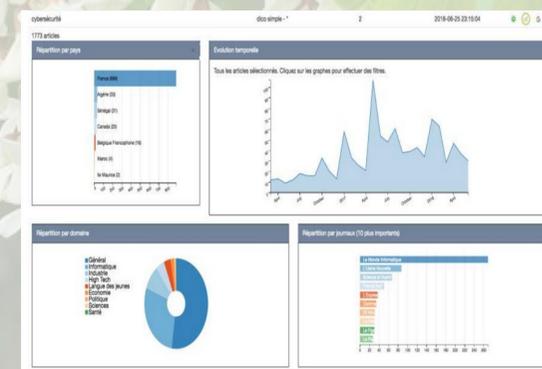
✓ Accès à plusieurs éléments pour l'analyse

✓ Itération sur processus de détection

- ❑ Injection des NC rejetés dans les lexiques d'exclusion

Néologisme candidat	Type	Commentaire	Reçu Automatique	Fréquence	Date	
abandon-éducation	des composé *		1	1	2018-09-23 16:33	🟢 🟡 🔴
fi-activité	des composé *		1	1	2018-09-23 16:17	🟢 🟡 🔴
super-yacht	Acture suggestion		1	1	2018-09-23 16:16	🟢 🟡 🔴
ultra-croquets	des composé *		1	1	2018-09-23 16:07	🟢 🟡 🔴
nettoyage-meur	des composé *		1	1	2018-09-23 16:05	🟢 🟡 🔴
sublim-eyes	des composé *		1	1	2018-09-23 16:05	🟢 🟡 🔴
pre-instruction	des composé *		1	1	2018-09-23 16:01	🟢 🟡 🔴
impregnation-éducation	des composé *		2	1	2018-09-23 15:59	🟢 🟡 🔴
se-pais	des composé *		1	1	2018-09-23 15:51	🟢 🟡 🔴
lunch-ritual	des composé *		1	1	2018-09-23 15:45	🟢 🟡 🔴
abstulude	Acture suggestion		1	1	2018-09-23 15:38	🟢 🟡 🔴

Le NFC en contexte et plusieurs informations complémentaires



Intérêts du processus itératif pour la détection

✓ Quelques chiffres

- ❑ *Mots inconnus seulement = 45 % de précision*
- ❑ *Avec réinjection des candidats rejetés pendant 2 semaines de travail = 60 % de précision*
- ❑ *Jusqu'à 90 % avec des réseaux de neurones (expériences en cours)*

Protocole de validation des NC

- ✓ Réduire le recours au « sentiment néologique »
 - ❑ *Analyse et classification des non-néologismes*
 - ❑ *Pas d'attestation avant 2010*
 - ❑ *Recours aux ressources existantes : corpus, dictionnaires de référence, dictionnaires historiques*
- ✓ Validation finale par vote majoritaire entre experts
 - ❑ *Une réunion mensuelle depuis 2015*
 - ❑ *Juin 2015 – fin 2017 : pour le français, validation de 21000 néologismes*

Fiches d'attestation des néologismes validés (1)

- ✓ Propriétés linguistiques indiquées par les experts
 - *POS, classe sémantique (Le Pesant et Mathieu-Colas 1998), le ou les procédés néologiques impliqués (Sablayrolles 2016)*
 - *Configuration syllabique et construction morphologique*
 - *Vocable(s) de base et POS vocable(s) de bas*
- ✓ Propriétés linguistiques récupérées par la plateforme
 - *Famille morphologique du néologisme*
 - *Profil combinatoire (collocations, constructions les plus fréquentes)*
 - *Profil distributionnel (lexies sémantiquement proches)*
 - × (quasi-)synonymes, hyperonymes, hyponymes

Fiches d'attestation des néologismes validés (2)

✓ Propriétés socio-pragmatiques

- ❑ *Public visé : audience générale ou public ciblé (ex. presse féminine)*
- ❑ *Type de document*
 - × Uniquement des articles de presse en 2019
- ❑ *Domaine : issu de l'information thématique fournie par les auteurs des documents où apparaît le néologisme validé*
- ❑ *Pays ou région d'origine des journaux contenant les documents*

Synthèse des néologismes validés en 2019

Mécanisme néologique principal	Nombre de néologismes (formes uniques)		Nombre d'occurrences de néologismes		Moyenne d'occ. par forme néologique
	Nombre	%	Nombre	%	
préfixation	17 051	75,87 %	485 566	66,86 %	28
composition	1 646	7,32 %	31 173	4,29 %	19
emprunt	1 429	6,36 %	132 104	18,19 %	92
suffixation	1 245	5,54 %	65 262	8,99 %	52
fracto-composition	791	3,52 %	7 039	0,97 %	9
onomatopée	92	0,41 %	665	0,09 %	7
troncation	73	0,32 %	2 678	0,37 %	37
composition savante	68	0,30 %	479	0,07 %	7
compoction	47	0,21 %	1 043	0,14 %	22
composition hybride	33	0,15 %	213	0,03 %	6
mot-valise	9	0,04 %	100	0,01 %	11
Totaux	22 475	100,00 %	726 222	100,00 %	

Le Logoscope : accès

✓ Interface en ligne accessible

☐ <https://logoscope.unistra.fr/>



The screenshot shows the website interface for Logoscope. At the top, there is a navigation bar with a home icon and several menu items: Nouveautés, Ordre alphabét., Fréquence, Chronologie, JournAI, Catégorie gram., Thème, Procédé, and Position. A search button labeled 'Rechercher' is on the right. The main content area features the word 'Logoscope 2' in a large, stylized font, where the 'o' is replaced by a magnifying glass icon. Below this, the text reads: 'Documentation quotidienne des nouveaux mots français', '1493 néologismes', and 'Dernière mise à jour - août 2018'. On the left side, there is a vertical list of links: 'Projet et équipe', 'Version experte', 'Guide d'utilisation', 'Archive des hapax', 'Publications', and 'Mentions légales'. At the bottom, there are logos for 'lilpa' (linguistique, langues, parole), 'UNIVERSITÉ DE STRASBOURG', 'FranceTerme', and Creative Commons BY-NC-SA.

Le logoscope en ligne

✓ Un état des lieux synthétique de la ressource

Documentation quotidienne des nouveaux mots français

1493 néologismes

Dernière mise à jour - août 2018

✓ Documentations et informations

[Projet et équipe](#)

[Version experte](#)

[Guide d'utilisation](#)

[Archive des hapax](#)

[Publications](#)

[Mentions légales](#)

✓ Une barre de recherche multi-axes

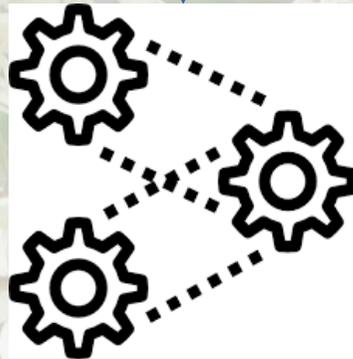
[Nouveautés](#) [Ordre alphabét.](#) [Fréquence](#) [Chronologie](#) [JournAl](#) [Catégorie gram.](#) [Thème](#) [Procédé](#) [Position](#)

Rechercher

Logoscope : structure



Aspiration automatisée
des sources



Détection NC

inextremiste Graphique 1

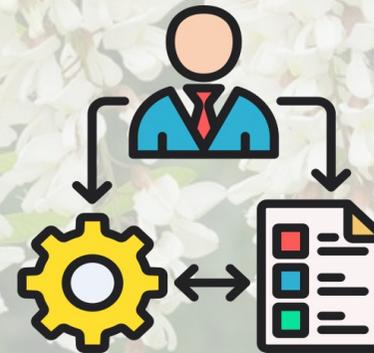
- Adjectif - CULTURE-Cinéma et Théâtre | CULTURE-Arts Plastiques et Photographie | CULTURE-Musique - Position initiale - Morphosémantique - 27/12/2015 - Les Echos - MARIANNE BLIMAN - Lien -

Contexte : "Les cirassiens de Cirque **inextremiste** n'ont pas encore fini avec « Extension » qu'ils travaillent déjà sur deux autres créations - dont un **spectacle** de rue de grand envergure avec une montgolfière. Un **processus** au long cours pour ces **artistes** engagés notamment pour « l'éducation populaire » et dont la **compagnie** fonctionne en auto-production, sans subvention. « Extension » a ainsi nécessité deux années de **création**, au cours desquelles le **spectacle** naissant a été montré à plusieurs **reprises** aux **spectateurs**. « J'ai l'impression d'un **verre** qui se remplit petit à petit »"

- Adjectif - CULTURE-Cinéma et Théâtre | CULTURE-Musique | LOISIRS-Sport - Position finale - Morphosémantique - 28/12/2015 - Les Echos - PHILIPPE CHEVILLEY - Lien -

Contexte : "« VIDEO Mené à un **tempo** d'enfer et joyeusement déjanté, le **dernier spectacle** du Cirque **inextremiste** traite du handicap et des relations humaines. A. "

BD de NC + documents annotés
thématiquement



Évaluation semi-automatique des NC
- classification statistique par
apprentissage (supervisé)
- exploitation d'une analyse thématique
probabiliste

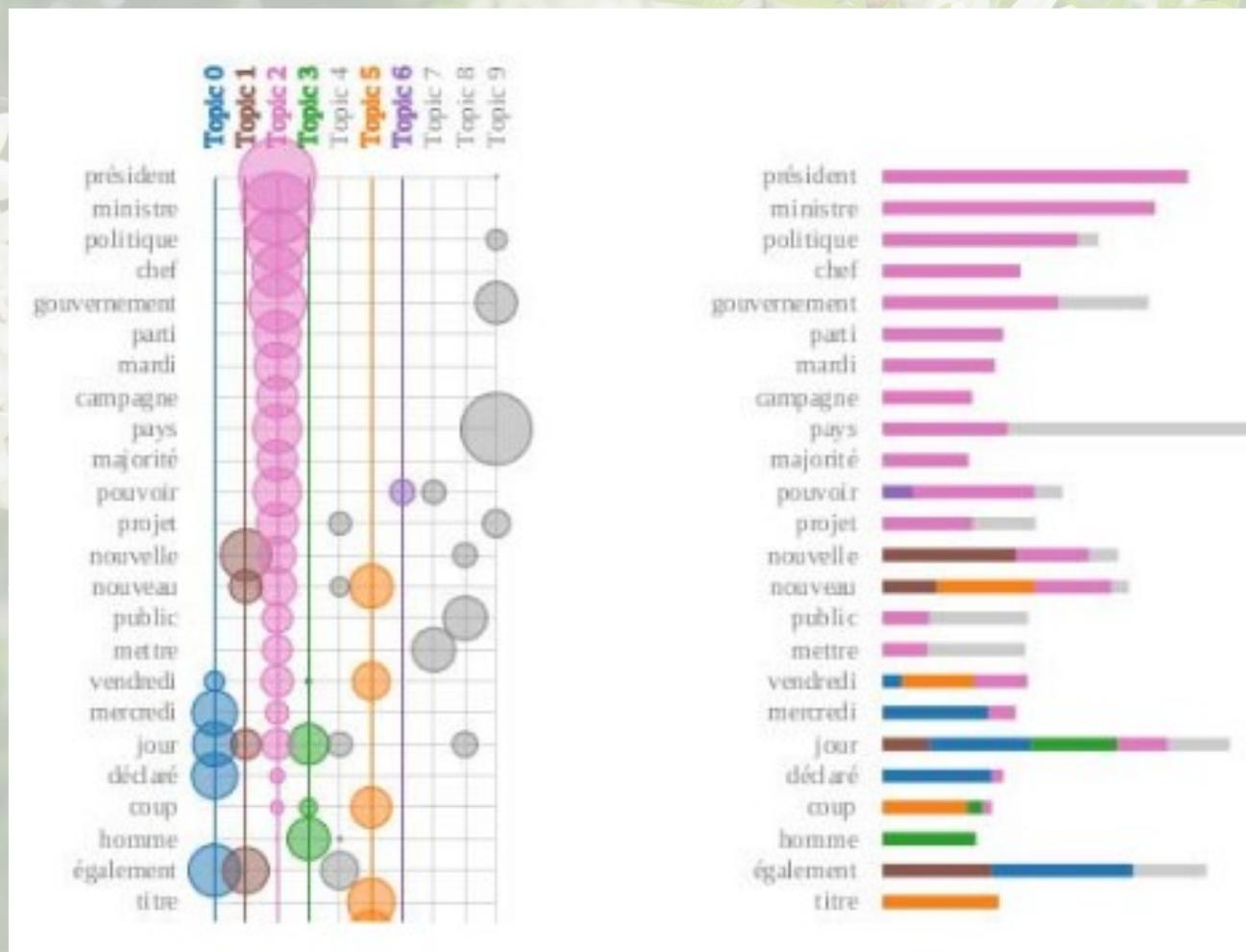
Logoscope : analyse thématique (1)

- ✓ Identification des thèmes de l'ensemble des sources
 - ❑ *Extraction d'une collection cohérente avec les sources des NC = 4755 documents pour les 10 sources*
 - ❑ *Extraction des vocables supposés pertinents pour la détection automatique de thèmes = ~7900 « mots pleins »*
- ✓ Mallet (Mc Callum 2002) avec 15 thèmes cibles et visualisation avec Termite (Chuang, Manning & Heer 2012)
 - ❑ *Distribution des mots pleins sur les thèmes cibles*

Logoscope : analyse thématique (2)

✓ Mallet + Termite

□ Exemple sur 10 thèmes cibles



□ MAIS Qualification manuelle de chaque thème via un « mot » sensé résumer la liste des mots les plus fréquents

Logoscope

✓ Exemple de résultat avec l'adjectif pluriel *climaticides*

climaticides [Graphique](#) [Wiktionnaire](#)

1

- Adjectif - [ÉCONOMIE-Finance](#) | [DROIT-Justice et Législation](#) | [ÉCONOMIE-Commerce](#) - Position finale - Morphosémantique - 21/06/2015 - Libération - CHRISTIAN LOSSON - [Lien](#) -

Contexte : "Alix Mazounie : « C' est un message important à envoyer à l' [Union Européenne](#) : globalement , le [Plan Juncker](#) n' est pas cohérent et compatible avec une politique climatique ambitieuse de l' Europe . L' efficacité énergétique reste le parent [pauvre](#) et le [plan](#) propose d' investir dans de nombreux [projets](#) d' infrastructures [climaticides](#) . Et l' [accord commercial](#) transatlantique va venir affaiblir plus encore les efforts fournis contre les changements climatiques . Tant que l' Europe ne fera pas de la transition énergétique l' objectif , le moyen et le [résultat](#) prioritaire , elle fait fausse route et le climat aussi . » "

- Adjectif - [ÉCONOMIE-Finance](#) | [DROIT-Justice et Législation](#) | [ÉCONOMIE-Commerce](#) - Position médiale - Morphosémantique - 21/06/2015 - Libération - CHRISTIAN LOSSON - [Lien](#) -

Contexte : "de [projets](#) incompatibles avec la lutte contre le changement climatique . Or , c' est trop souvent le [cas](#) et souvent au nom de la lutte contre la pauvreté - il faut impérativement inverser cette logique qui consisterait « au nom du [développement](#) » à , par exemple , renforcer la dépendance au charbon dans les pays les plus [pauvres](#) au lieu de leur donner un accès aux énergies renouvelables et locales . Cette initiative devrait également s' [appliquer](#) au Fonds vert pour le climat qui ne s' est pas encore fixé comme règle de ne pas [financer](#) de [projets](#) [climaticides](#) . On marche sur la tête . Dans tous les [cas](#) , il faut affiner la définition de " co-bénéfice " pour le climat qui reste encore trop vague et peut recouvrir un peu tout et n' importe quoi . » "

- Adjectif - [ÉCONOMIE-Finance](#) | [ÉCONOMIE-Commerce](#) | [ÉCONOMIE-Industrie](#) - Position finale - Morphosémantique - 06/11/2015 - Les Echos - JEAN MICHEL GRADT - [Lien](#) -

Contexte : "« BNP Paribas n' a décidément pas peur du ' greenwashing ' : elle ose s' afficher comme grande mécène de la COP21 alors qu' elle est la pire [banque](#) française en [matière](#) de [financements](#) [climaticides](#) , et qu' elle n' a pris cette année aucun engagement pour [réduire](#) ses soutiens aux énergies fossiles , contrairement au [Crédit Agricole](#) et à Natixis " , souligne Lucie Pinson , chargée de campagne [Finance](#) privée/Coface chez Amis de la Terre France . "

Adjectif - [ÉCONOMIE-Finance](#) | [ÉCONOMIE-Commerce](#) | [ÉCONOMIE-Industrie](#) - Position médiale - Morphosémantique - 06/11/2015 - Les Echos - JEAN MICHEL GRADT - [Lien](#) -

Logoscope vs. Néoveille

	Logoscope	Néoveille
Langues	Français	Chinois, Français, Grec, Polonais, Tchèque, Portugais du Brésil, Russe
Sources journalistiques	10	250
Période	2014 - 2018	2015 - maintenant
Accessibilité	Plateforme en ligne avérée Version experte en dormance	Plateforme en ligne théorique Version github licence Apache 2.0 à explorer
Nombre de Néologismes	1483	> 20.000
Validation NC	Automatique supervisée	Aide à la décision + groupe d'experts
Fiches d'attestation	Non	Oui
Caractérisation thématique	Analyse automatique probabiliste	Extraction des métadonnées auteurs
Base textuelle et néologique annotée	Oui	Oui