Data & Corpus, 2 | 2026
Nouveaux enjeux sociétaux et innovations méthodologiques

Article de données

Corpus multilingues pour l'étude de nouveaux concepts en SHS : le cas de « l'innovation non technologique »

KYRIAKOGLOU Revekka * [1], PAPPA Anna, [1]

[1] UR 4383 LIASD, Université Paris 8
* Mail : kyriakoglou@up8.edu

Domaines
Sciences humaines et sociales — Linguistique

Informatique — Algorithme et structure de données

Date de soumission : 8 décembre 2025
Date d'acceptation : 8 juin 2026
Date de publication : 23 juillet 2026

© The Authors
Licence : Creative Commons Attribution International (CC BY-NC 4.0)
Publié par l'Université de Lorraine

DOI : https://doi.org/10.46298/dc.17050

Résumé
Cet article présente une méthodologie hybride pour la création d'un corpus multilingue destiné à faire émerger des termes associés à des concepts nouveaux en sciences humaines et sociales (SHS), en prenant pour cas d'étude « l'innovation non technologique ». Le corpus repose sur deux sources textuelles complémentaires : (1) des contenus extraits automatiquement des sites web d'entreprises, nettoyés pour le français et l'anglais, et (2) des rapports annuels collectés et filtrés automatiquement selon des critères documentaires (année, format, duplication). Le pipeline intègre la détection automatique des langues, le filtrage des contenus non pertinents, l'extraction de segments pertinents et l'enrichissement en métadonnées structurelles.

À partir de ce corpus initial, des *datasets* dérivés sont construits pour des usages en analyse exploratoire et en apprentissage automatique : ils regroupent des représentations vectorielles et des métadonnées non textuelles, ainsi qu'un sous-ensemble structuré en blocs contextuels contenant des occurrences de termes annotés selon leur catégorie thématique.

Cette approche produit une ressource reproductible, extensible et adaptée à l'analyse de la variabilité lexicale autour de concepts nouveaux, ainsi qu'à la création de jeux de données destinés aux applications de traitement automatique des langues.

Mots-clés corpus, multilingue, traitement automatique des langues naturelles, extraction de termes, représentation vectorielle, apprentissage automatique, innovation non-technologique

Abstract
This article presents a hybrid methodology for building a multilingual corpus designed to support the study of emerging concepts in the humanities and social sciences (HSS), illustrated here through the case of "non-technological innovation". The corpus relies on two complementary sources: (1) textual content automatically extracted from company websites, cleaned for French and English, and (2) annual reports collected and automatically filtered according to documentary criteria (year, format, duplication). The processing pipeline includes automatic language detection, filtering of non-relevant content, extraction of relevant segments, and enrichment with structural metadata.

From this initial corpus, derived datasets are created for exploratory analysis and machine learning purposes. They include vector representations, non-textual metadata, and a subset organized into five-sentence contextual blocks centered on occurrences of terms from the expert lexicon. Each occurrence is annotated with its associated thematic category.

This approach results in a reproducible and extensible resource, suitable both for analyzing lexical variability around emerging concepts and for generating datasets dedicated to natural language processing applications.

Keywords corpus, multilingual, natural language processing, term extraction, vector representation, machine learning, non-technological innovation

Informations sur le corpus

Titre	MALANTIN (MAchine Learning for Analysing Non-Technological INnovation)
Auteures	Kyriakoglou, Revekka – Université Paris 8 Pappa, Anna – Université Paris 8
Date de publication	2025-12-05
Version	V1.
Lien	https://code.up8.edu/glacon/malantin_dataset_release/-/tree/main https://doi.org/10.57745/M9UHMX
Identifiant unique	https://doi.org/10.57745/M9UHMX
Disponibilité	Les données dérivées diffusées avec cet article sont disponibles dans le dépôt GitLab suivant : https://code.up8.edu/glacon/malantin_dataset_release/-/tree/main Le dépôt contient les représentations vectorielles non directement réversibles, les métadonnées non textuelles, le fichier `README.md`, le fichier `LICENSE.txt` et le fichier `citation.bib` Une copie des données est également disponible via Ortolang : https://repository.ortolang.fr/api/content/malantin/head/

Licence d'utilisation	Open Data Commons Attribution License (ODC-BY 1.0).
Formats et métriques	• Collection pages web + PDF (bilans entreprises) : **Pages web multilingues** : collectées automatiquement, filtrage d'URL, extraction ciblée des balises textuelles, détection de la langue et enrichissement en métadonnées (secteur, source, mots-clés, longueur). **Rapports annuels PDF** : collectés via recherche dirigée, extraction textuelle automatisée, normalisation, déduplication et contrôle qualité. **Non redistribués** : les contenus web et PDF ne sont pas publiés ; seules des représentations vectorielles et des métadonnées non textuelles sont diffusées pour garantir la protection des données. • Données vectorielles au format *.npy* et métadonnées au format *.csv* • Nombre, ampleur, poids des fichiers : Dataset anglais (en) : 3 fichiers, 1,3 MB Dataset français (fr) : 3 fichiers, 98 KB Dataset machine learning (ml) : 3 fichiers, 38,6 MB • Nommage des fichiers : Pour les jeux de données français et anglais, les fichiers vectoriels suivent le schéma `<langue>_embeddings_<méthode>_<dimension>.npy` et les métadonnées le schéma `<langue>_metadata_no_text.csv` Pour le *dataset machine learning*, les fichiers suivent le préfixe ml_sentence_ Les noms indiquent *: la langue, la méthode de réduction (pca / umap), la dimension finale, l'absence de texte.* • Arborescence des dossiers et les règles : *malantin_dataset_release/* *├── README.md* *├── LICENSE.txt* *├── citation.bib* *└── dataset_release/* *├── french/* *│ ├── fr_embeddings_pca_117.npy* *│ ├── fr_embeddings_umap_64.npy* *│ └── fr_metadata_no_text.csv* *├── english/*

	│ ├── *en_embeddings_pca_128.npy* │ ├── *en_embeddings_umap_64.npy* │ └── *en_metadata_no_text.csv* └── *machine_learning_dataset/* ├── *ml_sentence_embeddings_pca_128.npy* ├── *ml_sentence_embeddings_umap_64.npy* └── *ml_sentence_metadata_no_text.csv* Le dossier dataset_release/ correspond au jeu de données dérivé mis à disposition avec l'article. Il contient uniquement des représentations vectorielles non directement réversibles au format .npy et des métadonnées non textuelles au format .csv. Les textes bruts issus des pages web et des rapports annuels ne sont pas redistribués. Le fichier `README.md` précise la structure du dépôt, les usages possibles des fichiers et le lien entre les données diffusées et la méthodologie décrite dans l'article.
Langue	fr, en
Publication liée	https://hal.science/hal-04772221v1
Référence bibliographique	Kyriakoglou, R., Pappa, A., He, J., Schoen, A., Laurens, P., Vartampetian, M., Laredo, P. & Kyriakopoulou, T. (2024). Développement automatique de lexiques pour les concepts émergents : une exploration méthodologique. Dans A. Dister & D. Longrée (dir.), *JADT 2024 : Mots comptés, textes déchiffrés* (vol. 1, Cahiers du CENTAL, n° 17). Presses universitaires de Louvain. hal-04772221

1. Introduction

L'utilisation de corpus spécifiques constitue un enjeu central pour l'analyse des concepts émergents en sciences humaines et sociales (SHS), en particulier lorsqu'il s'agit des notions encore peu couvertes ou insuffisamment représentées dans les ressources lexicales existantes. Plusieurs structurations conceptuelles dérivées de *Princeton WordNet*[1] ont été proposées, comme *WordNet Domains* qui ajoute une hiérarchie thématique aux *synsets* (Bentivogli et al., 2004) puis son extension

[1] https://wordnet.princeton.edu/

multilingue révisée (Gonzalez-Agirre et al., 2012). D'autres ressources de large couverture, telles que *BabelNet, qui combine WordNet et Wikipedia* (Navigli & Ponzetto, 2012), *Open Multilingual WordNet* (Bond et al., 2013) et *ConceptNet* (Speer et al., 2017) offrent des représentations conceptuelles utilisées pour l'analyse textuelle. Toutefois, ces ressources, centrées sur des concepts stabilisés et encyclopédiques, se révèlent moins adaptées à l'étude de notions émergentes ou interdisciplinaires, qui nécessitent des corpus thématiques contextualisés et dynamiques.

L'exploration de ces concepts requiert la constitution de corpus spécialisés permettant d'identifier, de contextualiser et d'analyser les usages lexicaux qui leur sont associés (Kyriakoglou et al., 2024). Ces corpus ne servent pas seulement à enrichir les ressources existantes : ils fournissent une base empirique pour la création de jeux de données destinés à l'analyse linguistique, comparative ou socio-économique. Dans le cadre de cette étude consacrée à « l'innovation non technologique », l'objectif est de construire une ressource multilingue permettant d'observer et de comparer les variations lexicales à travers différents secteurs d'activité et différents contextes de communication organisationnelle.

Ce travail s'inscrit dans le projet MALANTIN, consacré à l'analyse de données textuelles pour l'identification de concepts émergents, avec pour cas d'étude l'innovation non technologique. Le projet vise à développer une méthodologie combinant la constitution de corpus thématiques, l'expertise humaine, l'analyse statistique et l'apprentissage automatique afin d'identifier, de structurer et d'enrichir les termes associés à ce domaine. Dans les travaux antérieurs liés au projet, cette approche a notamment permis de constituer un corpus multilingue à partir de sites web d'entreprises et de rapports annuels, puis d'élaborer un lexique de référence et des jeux de données annotés destinés à l'entraînement de modèles d'identification de nouveaux termes. Le présent article prolonge ce cadre en se concentrant plus spécifiquement sur la construction, la documentation et la mise à disposition contrôlée du corpus et des données dérivées.

Dans ce cadre, le protocole s'appuie sur une liste initiale de 3 992 entreprises provenant de la base CIB, utilisée comme point d'entrée pour la recherche et la collecte automatisées. Toutefois, seules les entreprises dont les sites ou rapports répondaient aux critères de qualité, d'accessibilité et de filtrage ont pu être effectivement intégrées au corpus final (Pappa & Kyriakoglou, 2025). Le corpus proposé combine l'extraction automatisée, la détection et la vérification de la langue, le filtrage des contenus non pertinents et l'enrichissement en métadonnées. Il repose sur deux ensembles complémentaires : (1) des contenus textuels extraits automatiquement des sites web institutionnels des entreprises effectivement accessibles, couvrant 27 secteurs économiques ; (2) des rapports annuels publiés entre 2017 et 2021, collectés par extraction automatisée, normalisation et filtrage documentaire. Ce périmètre assure une couverture spatio-temporelle cohérente, adaptée à l'analyse de phénomènes lexicaux contemporains et à l'étude comparative de contextes sectoriels. À partir de ce corpus initial, des *datasets* dérivés sont construits pour des usages en analyse exploratoire et

en apprentissage automatique : ils regroupent des représentations vectorielles non directement réversibles et des métadonnées non textuelles, ainsi qu'un sous-ensemble structuré en blocs contextuels contenant des occurrences de termes annotés selon leur catégorie thématique.

La section suivante présente les travaux relatifs à ce domaine et situe la méthodologie proposée par rapport aux recherches existantes sur les corpus multilingues et l'analyse de concepts émergents.

2. État de l'art

Les corpus thématiques constituent un outil essentiel pour l'analyse des représentations linguistiques, culturelles et sociales. Au-delà des ressources lexicales générales évoquées dans l'introduction, plusieurs travaux ont exploré la création de corpus spécialisés permettant de documenter la diversité des usages linguistiques dans des environnements variés et de soutenir l'étude de notions complexes ou en évolution.

La montée en puissance des corpus issus du web a notamment joué un rôle structurant. Plusieurs travaux ont montré que l'extraction massive de textes en ligne permettait de constituer des corpus étendus et représentatifs (Baroni et al., 2009, Kilgarriff et al., 2010). Dans cette perspective, des études telles que (Biemann et al., 2013) ou (Nakov, 2015) ont confirmé que les données web offraient un terrain privilégié pour observer la variabilité lexicale et sémantique dans des contextes réels. De plus, les travaux sur l'identification automatique de la langue ont renforcé la qualité des pipelines de collecte et de prétraitement des données textuelles (Lui et al., 2012).

Pour aller au-delà des ressources généralistes, plusieurs initiatives récentes se sont orientées vers des corpus thématiques enrichis d'annotations structurées. Le projet MOSAICo (Conia et al., 2024) relie le texte brut à des bases de connaissances pour permettre une annotation interconnectée. Cette approche s'inscrit dans la continuité des travaux méthodologiques décrits dans le *Handbook of Linguistic Annotation* (Ide et al., 2017). D'autres travaux, comme AlphaMWE (Han et al., 2020), enrichissent les corpus parallèles par des annotations lexicales fines, portant sur les expressions polylexicales.

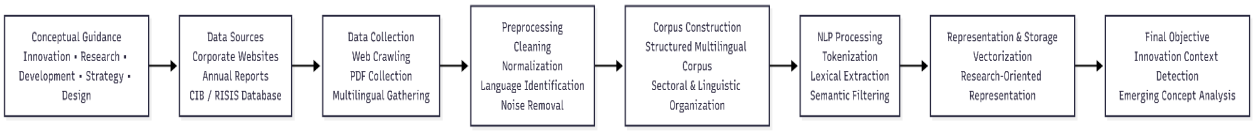
Enfin, des infrastructures telles que SDHSS (Beretta, 2024) qui s'appuie sur des modèles conceptuels FAIR comme CIDOC CRM, ou la plateforme européenne CLARIN (Hinrichs & Krauwer, 2014) renforcent la diffusion, la documentation, l'interopérabilité et la durabilité des données dans des contextes interdisciplinaires.

Ces travaux marquent une évolution vers des corpus sémantiquement enrichis, mieux intégrés dans les écosystèmes de données des SHS. Cependant, la question des concepts émergents, souvent flous, évolutifs et encore dépourvus de formes lexicales stabilisées, demeure largement ouverte. Peu de méthodes existantes proposent un pipeline reproductible permettant de les détecter, de les documenter et de les structurer à partir de corpus multilingues contextualisés.

La méthodologie présentée dans cet article s'inscrit dans cette perspective : elle propose une approche complète, allant de la collecte automatisée à la structuration sémantique, pour construire et exploiter des corpus thématiques destinés à l'identification et à l'étude de concepts émergents.

3. Description de la méthodologie

Figure 1. Vue d'ensemble du pipeline méthodologique



Note : Le processus comprend la collecte des données (sites web et rapports annuels), le filtrage et l'enrichissement des contenus, la structuration des sous-corpus, la génération de représentations vectorielles, puis la construction de ressources lexicales et de jeux de données dérivés destinés à l'analyse et à l'apprentissage automatique.

3.1. Cadre général et sources de données

La méthodologie adoptée pour la construction du corpus repose sur une approche hybride et systématique, combinant extraction automatisée, filtrage sémantique, structuration multilingue et vectorisation des contenus. Elle vise à constituer une ressource homogène, représentative des contenus textuels décrivant les activités des entreprises, et à offrir un espace d'observation pour la détection des contextes liés à l'innovation. Le corpus repose sur deux types de documents complémentaires : les contenus textuels extraits des sites web et les rapports annuels des entreprises.

Cinq mots-clés conceptuels (*innovation*, *recherche*, *développement*, *stratégie* et *design*) ont servi de guide à l'ensemble du processus, depuis la sélection des sources jusqu'à l'extraction lexicale. Les entreprises ciblées proviennent de la base de données CIB (*Corporate Innovation Base*) du projet européen RISIS 2 (*Research Infrastructure for Science and Innovation Policy Studies*), qui recense les principales firmes européennes impliquées dans des activités d'innovation et de R&D (Recherche et Développement). La base CIB fournit un échantillon d'entreprises impliquées dans des activités d'innovation, sélectionnées selon des critères harmonisés au niveau international. Elle offre un point d'entrée structuré pour identifier des acteurs industriels pertinents dans divers domaines technologiques et économiques.

Au total, 3 992 grandes entreprises, réparties dans 27 secteurs économiques, ont été retenues et validées par les experts du projet.

Cette base constitue le jeu de référence pour l'exploration des sites web et la collecte des rapports annuels, garantissant la cohérence sectorielle et linguistique du corpus. Les données collectées sont exclusivement exploitées à des fins de recherche scientifique et transformées en représentations vectorielles pour garantir la conformité éthique et juridique (Kazmali & Sayar, 2025) de leur diffusion.

3.2. Sous-corpus pdf

Le sous-corpus PDF regroupe les rapports annuels publiés par les entreprises de la base CIB. Ces documents décrivent les orientations stratégiques, les investissements technologiques ainsi que les activités économiques et organisationnelles, ce qui en fait une source adaptée pour analyser la mise en discours de l'innovation dans différents contextes industriels. Ils permettent d'identifier les usages lexicaux liés à l'innovation technologique ou non technologique et d'observer leurs variations selon les secteurs. La collecte a été réalisée à l'aide d'un analyseur automatisé développé sur la base de la bibliothèque *libcurl* (outil *open* source de gestion de requêtes HTTP), permettant le téléchargement en masse des rapports tout en appliquant un filtrage strict des URL selon des critères prédéfinis, fondés sur des motifs lexicaux et structurels (inclusion du domaine de l'entreprise, exclusion des URL associées à des réseaux sociaux ou à des sites commerciaux ou de divertissement).

Cette procédure repose sur un fichier dictionnaire associant le nom et le domaine de chaque entreprise, fourni par RISIS 2. Chaque entrée de ce fichier, formulée selon la syntaxe :

COMPANY + intext:"Annual Report" AND inurl:domain

a servi de requête automatisée via l'API de recherche Google ou par interrogation semi-automatisée, garantissant la localisation des rapports officiels. Les formats de traitement de texte (.doc, .rtf) ont été écartés en raison de leur hétérogénéité structurelle, afin de conserver uniquement les fichiers PDF standardisés et compatibles avec l'extraction automatisée.

Pour compléter la collecte, des rapports supplémentaires ont été récupérés depuis le site *AnnualReports.com* (https://www.annualreports.com), qui centralise les publications financières internationales.

La collecte a été réalisée dans le respect des politiques d'accès des sites (robots.txt) et sans contournement de dispositifs de protection. Seules les ressources librement accessibles ont été traitées conformément aux bonnes pratiques de *scraping* éthique et légal (Essien et al., 2023).

Les textes extraits ont fait l'objet d'un prétraitement comprenant :

- une normalisation typographique (normalisation des encodages et des caractères spéciaux suppression retours à la ligne, harmonisation des espaces),

- la tokenisation,
- un filtrage lexical minimal destiné à éliminer les segments dépourvus d'information linguistique utile.

Deux seuils de qualité ont été appliqués : exclusion des rapports antérieurs à 2017 et des documents contenant moins de 1 000 tokens après nettoyage, afin de garantir une densité informationnelle suffisante. Une procédure de détection automatique des doublons a permis de ne conserver qu'un rapport par entreprise et par année, en privilégiant le plus récent.

Au total, 8 368 rapports annuels publiés entre 2017 et 2021 ont été rassemblés, assurant une couverture sectorielle et temporelle cohérente avec le périmètre d'étude. Ces documents constituent un sous-corpus homogène et comparable, adaptés à l'analyse textuelle et à l'extraction terminologique. Ils serviront de base à la création de représentations vectorielles et à l'identification des contextes lexicaux présentés dans les sections suivantes.

3.3. Sous-corpus web

Le corpus web regroupe les contenus textuels publiés sur les sites des entreprises, décrivant leurs activités, leurs projets et leurs orientations stratégiques. Ces textes permettent d'identifier des segments susceptibles d'évoquer l'innovation, qu'elle soit technologique ou non technologique. Un ensemble de *scrapers Python*, développés à l'aide du framework *Scrapy* (Myers & McGuffee, 2015), a été déployé pour explorer automatiquement les domaines des 3 992 entreprises de la base CIB.

3.3.1. Filtrage et sélection des pages

Avant toute extraction textuelle, un filtrage systématique des URL a été appliqué afin de limiter le bruit informationnel et de se concentrer sur les pages présentant un contenu pertinent. Des expressions régulières ont permis d'exclure plusieurs catégories de pages non informatives ou non pertinentes pour l'analyse des activités et projets des entreprises :

- pages d'authentification ou de compte utilisateur (*login, register, password reset*),
- contenus institutionnels génériques (FAQ, *legal notice, cookie policy*),
- sections de recrutement (*careers, jobs, people*),
- pages techniques ou utilitaires (view, print, thumbnail). L'extraction a ciblé les balises HTML à forte valeur sémantique : **<p>, <title>, <h1>** et **<h2>**. Les pages contenant au moins un des cinq mots-clés conceptuels définis en amont ont été retenues comme points d'entrée pour la sélection des contextes pertinents.

3.3.2. Détection linguistique et enrichissement

La détection de la langue a été effectuée en deux étapes complémentaires. Dans un premier temps, l'information linguistique déclarée dans la balise HTML**<lang>** a été utilisée lorsqu'elle était disponible. Dans un second temps, une vérification automatique a été réalisée à l'aide de la bibliothèque *googletrans* (https://pypi.org/project/googletrans), afin de corriger les erreurs fréquentes de métadonnées et d'assurer une identification fiable.

Chaque document a ensuite été enrichi par un ensemble de métadonnées descriptives : nom de l'entreprise, URL, source, secteur d'activité, langue détectée et occurrences de mots-clés conceptuels. Les textes ont été exportés aux formats texte brut et CSV afin de faciliter les traitements et exploitations ultérieurs.

3.3.3. Couverture et diversité linguistique

La collecte a permis d'exploiter 924 sites web d'entreprises, couvrant 27 secteurs économiques et 16 langues principales, après application d'un seuil de qualité fixant à 1 000 le nombre minimal de pages accessibles par domaine. Bien que tous les domaines initialement identifiés n'aient pas pu être intégralement explorés, l'échantillon obtenu reste représentatif du périmètre visé, tant sur le plan sectoriel que linguistique. Les secteurs d'activité couverts sont les suivants :

Tableau 1. Secteurs d'activité des entreprises

Secteurs d'activité
Machines électriques, électronique industrielle
Chimie, pétrole, caoutchouc et plastique
Services aux entreprises
Fabrication d'équipements de transport
Communications
Métaux et produits métalliques
Logiciels informatiques
Commerce de gros
Produits alimentaires, fabrication de tabac
Banque, assurance et services financiers
Biotechnologie et sciences de la vie
Matériel informatique
Mines et extraction
Services publics
Produits en cuir, pierre, argile et verre
Fabrication de meubles
Construction
Médias et diffusion

Secteurs d'activité
Fabrication de textiles et de vêtements
Commerce de détail
Transport, fret et stockage
Fabrication diversifiée
Voyages, services personnels et loisirs
Administration publique
Impression et édition
Agriculture, horticulture et élevage
Services immobiliers

Cette diversité sectorielle et linguistique constitue une base solide pour les étapes de traitement automatisé et de vectorisation décrites dans les sections suivantes.

3.4. Pipeline technique et algorithme

Le pipeline technique repose sur une architecture modulaire combinant l'extraction automatisée, la détection linguistique, le filtrage lexical fondé sur les mots-clés conceptuels, la structuration des contenus textuels et, dans une étape ultérieure du pipeline, la production de représentations vectorielles non directement réversibles. Cette organisation garantit une chaîne de traitement cohérente, reproductible et conforme aux exigences éthiques.

3.4.1. Extraction automatisée et structuration du corpus

Les documents collectés (pages web et rapports pdf) sont intégrés dans un pipeline automatisé comprenant quatre opérations principales :

- analyse HTML,
- détection linguistique,
- filtrage du contenu en fonction de la présence de mots-clés conceptuels,
- extraction et agrégation des données.

Chaque document textuel collecté est intégré dans ce pipeline, analysé et nettoyé, enrichi de métadonnées, puis agrégé dans les fichiers de sortie. Le schéma algorithmique suivant décrit la procédure complète de collecte et de construction du corpus.

Algorithme 1. Procédure de construction automatisée du corpus multilingue

```
Entrée : liste de domaines d'entreprises D, lexiques {keywords[lang]}
Sorties : corpus multilingue C, fichiers CSV de synthèse, (optionnel) liste de liens PDF identifiés
procedure BuildCorpus(D)
    C ← ∅
    for all domain ∈ D do
        assert is_allowed_by_robots(domain)
        html_pages ← crawl(domain, rate_limit=polite)
        html_pages ← apply_url_filters(html_pages)
        for all page ∈ html_pages do
            text_raw ← extract_visible_text(page, tags={p,title,h1,h2})
            text ← normalize_whitespace(text_raw)
            if length_in_tokens(text) < MIN_TOKENS then continue
            lang_html ← detect_from_html_attribute(page)
            lang_auto ← detect_language(text)
            lang ← validate(lang_html, lang_auto)
            K ← keywords[lang]
            if contains_any(text, K) then
                counts ← count_occurrences(text, expand_variants(K, lang))
                snippet_set ← extract_context_snippets(text, K, window=W)
                snippet_set ← deduplicate_snippets(snippet_set)
                meta ← {domain, url=page.url, lang, sector(domain), token_count, date_seen=now()}
                C ← C ∪ build_records(snippet_set, counts, meta)
            end if
        pdf_links ← find_pdf_links(page)
        pdf_links ← apply_url_filters(pdf_links)
        for all link ∈ pdf_links do
            if is_annual_report(link) then
                enqueue_pdf_for_pipeline_3_2(link, domain)
```

```
                    end if
                end for
            end for
        end for
        C ← deduplicate_records(C, keys={domain,url,snippet_hash})
        aggregate_results_to_csv(C)
        return C
    end procedure
```

Cette architecture modulaire permet une adaptation aisée à d'autres ensembles d'entreprises ou domaines thématiques. Elle garantit également la traçabilité des opérations, la reproductibilité du processus et la conformité éthique de la collecte, notamment vis-à-vis des règles d'accès aux contenus en ligne.

3.4.2. Vectorisation et production des données dérivées

En complément de la structuration textuelle, une représentation vectorielle non directement réversible a été produite afin de permettre la diffusion d'un jeu de données exploitable tout en respectant les contraintes légales liées aux droits d'auteur et aux données personnelles. Les phrases anonymisées, obtenues au moyen du pipeline Presidio assurant la reconnaissance automatique d'entités sensibles et leur substitution standardisée, constituent l'entrée du module de vectorisation.

Un tableau de données comprenant quatre champs (*word_labels, keywords, source, sentence_anonymized*) a servi de base au processus. Les phrases anonymisées ont été encodées à l'aide du modèle *paraphrase-multilingual-MiniLM-L12-v2* (*Sentence-Transformers*), générant pour chaque entrée un vecteur dense de dimension fixe.

Le choix du modèle paraphrase-multilingual-MiniLM-L12-v2 répond à la nécessité de disposer d'un encodeur multilingue compact, adapté à la représentation de phrases ou de blocs contextuels courts dans plusieurs langues. L'anonymisation préalable réduit certains indices lexicaux liés aux entités nommées, mais elle permet de limiter les risques associés à la diffusion de contenus sensibles ou protégés. La réduction dimensionnelle par ACP puis UMAP entraîne nécessairement une perte d'information par rapport aux *embeddings* initiaux ; elle est toutefois utilisée afin de produire des représentations plus compactes, plus facilement diffusables et exploitables pour des analyses exploratoires, des regroupements de contextes similaires ou des expérimentations de classification. Ces choix

impliquent que les vecteurs diffusés doivent être interprétés comme des représentations dérivées, non directement réversibles, adaptées à l'analyse computationnelle, et non comme des substituts complets des textes sources.

La réduction dimensionnelle a ensuite été appliquée en deux étapes :

- Une Analyse en Composantes Principales (ACP), avec un maximum de 128 composantes, sous la contrainte $n_{composantes} \leq (min\ n_{échantillons} - 1, n_{caractéristiques})$,
- une projection UMAP en 64 dimensions, adaptée aux usages exploratoires tels que le clustering, la visualisation ou la classification supervisée.

Les représentations finales sont distribuées sous forme de fichiers .npy, accompagnés de métadonnées non sensibles au format .csv. Dans le dépôt, ces fichiers sont organisés dans les sous-dossiers : `dataset_release/french/`, `dataset_release/english/` et `dataset_release/machine_learning_dataset/`. Les métadonnées associées incluent notamment les champs word_labels, keywords, source et, le cas échéant, doc_id. Aucun contenu textuel n'est publié : seuls les vecteurs non directement réversibles et les métadonnées non textuelles sont diffusés, garantissant la conformité éthique et juridique du *dataset* dérivé. Les jeux de données français, anglais et *machine learning* ont été traités selon le même pipeline de vectorisation.

3.5. Formalisation du processus : la démarche PRISME

Afin de garantir la transparence et la reproductibilité du processus, la méthodologie a été structurée selon une approche inspirée du cadre PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*; voir Page et al., 2021).

Adaptée à la constitution de corpus multilingues en SHS, cette approche est reformulée sous le sigle PRISME, qui désigne six étapes opérationnelles du pipeline : Préparation, Repérage, Identification, Structuration, Modélisation et Évaluation.

Chaque étape correspond à une phase clé du traitement, depuis la définition du périmètre conceptuel jusqu'à la validation du corpus final.

Tableau 2. Démarche PRISME pour la constitution du corpus

Étape	Objectif	Procédure	Outils principaux	Dépôt et données diffusées
Préparation	Définir le périmètre conceptuel et linguistique du corpus.	Sélection des concepts-clés (*innovation, recherche, développement, stratégie, design*) et constitution de la liste initiale d'entreprises.	Dictionnaire de mots-clés, base CIB (projet RISIS).	`README.md`.
Repérage	Identifier et collecter les sources pertinentes.	*Crawling* sélectif des domaines d'entreprises et recherche de rapports annuels.	Framework *Scrapy*, bibliothèque *libcurl*.	`KeywordsExtraction/spider/`.

Étape	Objectif	Procédure	Outils principaux	Dépôt et données diffusées
Identification	Extraire et segmenter les contenus textuels pertinents.	Analyse du DOM et extraction ciblée des balises `<p>`, `<title>`, *<h1>*, *<h2>*.	*XPath*, *BeautifulSoup*.	`KeywordsExtraction/pipelines/`
Structuration	Nettoyer et annoter les données textuelles.	Suppression du bruit, détection automatique de la langue, enrichissement par métadonnées.	*googletrans*, *langdetect*, *pandas*.	Les fichiers `.csv` de `dataset_release/`.
Modélisation	Organiser les sous-corpus et générer les fichiers structurés.	Production des sous-corpus web et PDF, exportation en fichiers CSV et texte brut, génération des représentations vectorielles non directement réversibles.	Scripts Python, pipeline d'analyse interne.	Les fichiers `.npy` sont diffusés dans `dataset_release/french/`, `dataset_release/english/` et `dataset_release/machine_learning_dataset/`.
Évaluation	Vérifier la qualité et la couverture du corpus.	Contrôle manuel d'un échantillon, analyse de la distribution lexicale et des occurrences par langue.	Analyse *n-grams,* validation interlinguistique.	-

Cette formalisation rend le processus de collecte transparent, reproductible et documenté, en précisant les outils, les paramètres et les critères appliqués à chaque étape. Afin de faciliter la réutilisation, chaque phase de la démarche PRISME est également associée aux éléments correspondants du dépôt : le code de collecte est documenté dans `KeywordsExtraction/spider/`, les opérations de nettoyage et de structuration dans `KeywordsExtraction/pipelines/`, et les représentations vectorielles non directement réversibles ainsi que les métadonnées non textuelles sont diffusées dans `dataset_release/`. Le respect des contraintes d'accès aux sites web (robots.txt, limitations de trafic) assure la conformité éthique de la collecte, tandis que la diffusion de fichiers standardisés (`.npy` et `.csv`) garantit la réutilisabilité du corpus pour des tâches d'analyse lexicale ou d'apprentissage automatique.

La section suivante présente l'implémentation concrète de cette méthodologie, depuis le prétraitement linguistique des textes collectés jusqu'à l'extraction et la structuration du lexique thématique.

3.6. Repérage et structuration des termes candidats

Cette section décrit la phase de repérage et d'organisation des termes-clés, conçus comme indicateurs lexicaux permettant d'identifier des contextes potentiellement associés à « l'innovation non technologique ». L'objectif n'est pas de constituer un lexique exhaustif, mais de définir un ensemble

structuré de termes de référence permettant de repérer les segments textuels où peuvent émerger de nouvelles formulations liées à la notion d'innovation. Cette phase combine analyse de surface (*n*-grammes) et pondération statistique (TF-IDF), afin d'identifier des termes représentatifs et fréquemment cooccurrents dans les textes d'entreprises.

3.6.1. Première étape : génération des *n*-grammes

Les textes collectés sont d'abord segmentés en *n*-grammes (séquences de *n* mots), avec *n* variant entre 2 et 5. Cette approche, indépendante des contraintes linguistiques spécifiques, permet de repérer des associations lexicales récurrentes y compris dans des langues et styles rédactionnels variés.

Un premier filtrage a été effectué pour écarter les *n*-grammes contenant des noms d'entreprises ou des séquences non lexicales (liens URL, adresses, éléments numériques, etc.).

Afin de concentrer l'analyse sur les domaines d'intérêt, seuls les *n*-grammes comportant au moins un mot appartenant à un ensemble de mots-cibles ont été conservés. Ces mots-cibles correspondent aux cinq concepts directeurs du projet (innovation, recherche, développement, stratégie, design) enrichis de deux termes additionnels (*new*, *product*), fréquemment associés aux contextes d'innovation selon les experts.

Enfin, un filtrage statistique a été appliqué afin de ne conserver que les *n*-grammes les plus significatifs, c'est-à-dire ceux dont la fréquence se situait au-delà de trois écarts types au-dessus de la médiane de distribution.

3.6.2. Deuxième étape : pondération *TF-IDF*

La seconde étape consiste à appliquer la pondération *TF-IDF* aux *n*-grammes issus du premier filtrage. Cette mesure statistique évalue l'importance d'un *n*-gramme dans un domaine d'entreprise au regard de sa fréquence dans l'ensemble du corpus. Chaque texte d'entreprise est traité comme un document unique, regroupant les contenus extraits des différentes URL d'un même domaine. Les textes sont ensuite comparés au sein de leur secteur d'activité, ce qui permet d'identifier les termes spécifiques à chaque domaine économique.

L'objectif est de faire émerger un lexique de référence fondé sur les régularités lexicales observées, plutôt que de partir d'un vocabulaire préétabli.

3.6.3. Illustration intersectorielle : pondérations comparées

Le tableau 3 présente, à titre d'exemple, les termes les mieux pondérés (selon la méthode TF-IDF) dans deux secteurs distincts, ainsi que leur intersection. Les valeurs indiquent le poids relatif de chaque terme dans l'ensemble des documents analysés, ce qui permet d'identifier les dénominations les plus caractéristiques du lexique de l'innovation dans chaque domaine.

Certains *n*-grammes, tels que *new product* ou *product development*, apparaissent plusieurs fois dans les deux secteurs en raison de leur fréquence élevée dans des contextes sémantiquement distincts.

Tableau 3. Intersection des termes les mieux pondérés par secteur

Services aux entreprises	**Poids du terme**	**Machines électriques et électroniques industrielles**	**Poids du terme**	**Intersection**
research development	1,00	research development	1,00	research development
sustainable development	0,73	products services	0,71	sustainable development
product development	0,65	product development	0,69	product development
product services	0,57	new product	0,45	product services
business development	0,54	sustainable development	0,43	business development
new product	0,49	business development	0,42	new product
innovative solution	0,46	innovative product	0,40	innovative solution
new technology	0,41	design development	0,38	new technology
new product	0,40	new product	0,38	new product
software development	0,37	innovative solution	0,37	innovative product
product solution	0,36	new technology	0,36	—
innovative product	0,35	clinical research	0,32	—
new business	0,35	product design	0,29	—
development process	0,35	development manufacturing	0,29	—
innovation award	0,34	new idea	0,26	—
development goal	0,34	quality product	0,25	—

Note : Les colonnes sont organisées en trois blocs indépendants : services aux entreprises, machines électriques et électroniques industrielles, et termes communs aux deux secteurs. Les lignes ne constituent pas des appariements terme à terme entre les blocs. Le symbole — signifie absence du terme dans l'intersection.

Un contrôle de redondance a également été effectué afin d'éliminer les paragraphes dupliqués, fréquents sur les sites web institutionnels. Les *n*-grammes présentant un score TF-IDF supérieur à un seuil défini sur la distribution des valeurs observées ont été retenus comme termes candidats pour le lexique final.

Cette double approche, linguistique (*n*-grammes) et statistique (TF-IDF), permet d'obtenir une sélection robuste de termes pertinents pour l'analyse thématique et la constitution de ressources lexicales sectorielles. Elle offre également une base exploitable pour l'apprentissage automatique, en facilitant la création de jeux de données annotés par secteur ou par domaine.

La section suivante présente les résultats issus de ce processus, en mettant en évidence la taille du corpus, la répartition sectorielle et linguistique ainsi que les premiers indices contextuels liés à l'innovation non technologique.

3.7. Construction du *dataset* dérivé en blocs de cinq phrases

Afin de produire un jeu de données exploitable pour des tâches d'apprentissage automatique, un sous-ensemble du corpus a été structuré sous forme de blocs contextuels centrés sur les occurrences des termes du lexique thématique. Pour chaque occurrence identifiée dans le corpus web ou PDF, un segment contextuel standardisé a été extrait, composé de cinq phrases consécutives : deux phrases précédant l'occurrence du terme ; la phrase contenant le terme ; deux phrases suivant cette occurrence.

Ce format rend possible l'apprentissage supervisé de modèles capables de prédire la catégorie thématique d'un terme à partir de son contexte immédiat. Chaque bloc est associé à un ensemble de métadonnées. Les blocs dépourvus de contexte suffisant (moins de cinq phrases disponibles) ont été conservés mais complétés par un padding standardisé.

Ces données sont par la suite anonymisées automatiquement à l'aide du pipeline Presidio et converties en représentations vectorielles non directement réversibles.

Ces choix méthodologiques ont des conséquences sur la couverture et la représentativité du corpus. Le seuil minimal de 1 000 tokens vise à exclure les documents trop courts ou peu informatifs, mais peut réduire la présence de certaines entreprises dont les sites web sont moins développés ou dont les rapports sont plus synthétiques. L'exclusion des formats non PDF permet de garantir une extraction textuelle plus homogène, au prix d'une couverture documentaire potentiellement plus restreinte. De même, la fenêtre contextuelle retenue pour la construction du *dataset* dérivé privilégie l'analyse du contexte immédiat des termes, mais ne permet pas de restituer l'ensemble de l'argumentation discursive des documents sources. Ces arbitrages visent à équilibrer qualité linguistique, homogénéité du traitement, conformité juridique et réutilisabilité des données diffusées.

4. Résultats

4.1. Validation de la méthode

Le corpus final couvre une grande variété de secteurs économiques et de langues, offrant une richesse lexicale autour des cinq mots-clés conceptuels. Les résultats présentés ici portent sur la constitution et la structuration du corpus, et non sur les modèles d'apprentissage automatique dérivés. L'analyse des données collectées montre une couverture sectorielle étendue, ainsi qu'une densité lexicale suffisante pour envisager des tâches ultérieures d'extraction ou de modélisation thématique.

Les expérimentations d'apprentissage automatique fondées sur ce *dataset* ont été présentées dans des travaux antérieurs liés au projet MALANTIN (Kyriakoglou et al., 2024). Dans le présent article, elles ne sont pas détaillées en tant que résultats de modélisation, mais mentionnées uniquement pour préciser les usages possibles du *dataset* dérivé et son potentiel de réutilisation.

Tableau 4. Répartition linguistique, sectorielle et volumétrique du corpus

Langue	Nombre d'URL	Secteurs couverts	Nombre de tokens
Anglais	41 224	27	8 133 370
Chinois	33 054	22	1 823 235
Français	15 074	17	2 821 719
Allemand	16 783	16	3 327 026
Espagnol	10 364	12	2 483 900
Italien	8 236	12	1 323 276
Coréen	10 929	11	1 899 763
Néerlandais	1 907	11	505 163
Catalan	2 427	9	1 149 180
Portugais	7 594	8	780 462
Danois	1 088	8	234 798
Russe	32 088	7	13 157 829
Suédois	1 355	7	219 159
Polonais	1 632	6	241 973
Finnois	1 029	4	99 361
Ukrainien	3 073	1	981

Le tableau 4 décrit la couverture linguistique du corpus initial collecté avant sélection, anonymisation et vectorisation. Il ne correspond donc pas directement aux jeux de données diffusés, qui sont limités aux sous-ensembles français et anglais ainsi qu'au *dataset* machine *learning* dérivé, accompagnés de métadonnées non textuelles.

La diversité linguistique observée reflète la variété lexicale et thématique attendue dans un corpus de grande ampleur couvrant plusieurs secteurs. Cette variété assure une meilleure représentativité des usages terminologiques et renforce la robustesse de la ressource pour des analyses comparatives et des expérimentations en traitement automatique du langage.

Le corpus ainsi obtenu constitue une ressource multilingue structurée, prête à être exploitée pour des analyses lexicométriques, l'extraction terminologique ou la création d'ensembles de données destinés à l'apprentissage supervisé. Dans cette perspective, un jeu de données dérivé sous forme de représentations vectorielles non directement réversibles a été produit, spécifiquement conçu pour les tâches d'apprentissage automatique.

À partir de ce corpus initial, des *datasets* dérivés sont construits pour l'analyse exploratoire et l'apprentissage automatique. Ils comprennent des représentations vectorielles non directement réversibles, des métadonnées non textuelles, ainsi qu'un sous-ensemble organisé en blocs contextuels de cinq phrases : la phrase contenant le terme, les deux phrases

précédentes et les deux phrases suivantes, soit environ 500 *tokens* en moyenne. Chaque occurrence est associée à une catégorie thématique.

Après vérification par les experts du projet, plus de 80 % des termes suggérés (dans un total de 700 termes proposés) ont été jugés pertinents et intégrés au lexique final, démontrant la fiabilité et l'efficacité du processus hybride combinant analyse automatique et validation humaine.

La validation experte a été réalisée par les membres du projet disposant d'une expertise métier en recherche et développement (R&D) et en analyse des dynamiques d'innovation. Les critères retenus concernaient la pertinence sémantique du terme proposé, sa cohérence avec la catégorie thématique associée et sa capacité à désigner un aspect de l'innovation non technologique. Cette validation humaine complète les contrôles automatiques de déduplication, de détection linguistique et de filtrage par seuils, afin de renforcer la fiabilité du lexique final et des données dérivées. Aucun calcul formel d'accord inter-annotateurs n'a été réalisé ; les divergences d'interprétation ont été traitées par discussion collective entre les expert·e·s jusqu'à stabilisation de la décision d'acceptation ou de rejet.

4.2. Réutilisation du *dataset* dérivé

Le *dataset* dérivé mis à disposition permet plusieurs formes de réutilisation, malgré l'absence volontaire des textes bruts. Les fichiers vectoriels au format `.npy` peuvent être exploités pour des tâches de clustering, de visualisation, de classification supervisée ou de comparaison sémantique entre secteurs et langues. Les fichiers de métadonnées au format `.csv` permettent d'associer chaque représentation vectorielle à des informations non textuelles, telles que la langue, le secteur, la source ou les catégories lexicales associées. Cette organisation rend le jeu de données directement exploitable pour des expérimentations en traitement automatique des langues, tout en respectant les contraintes juridiques et éthiques liées à la redistribution des contenus collectés. En revanche, l'absence de texte brut limite volontairement les usages fondés sur l'analyse qualitative fine ou la citation directe des documents sources.

Afin de faciliter l'appropriation de ces données par des chercheurs ne disposant pas nécessairement d'une expertise avancée en programmation, le dépôt fournit une documentation décrivant la structure des fichiers, les formats disponibles et les usages possibles du *dataset*. Les fichiers `.csv` peuvent être consultés directement dans un tableur afin d'examiner les métadonnées associées aux représentations vectorielles, tandis que les fichiers `.npy` peuvent être mobilisés, à l'aide de scripts simples, pour des analyses exploratoires telles que le regroupement de contextes similaires, la comparaison de secteurs d'activité ou l'entraînement de classifieurs supervisés.

4.3. Transposabilité et faisabilité du protocole

La méthodologie proposée peut être transposée à d'autres concepts émergents en SHS, à condition d'adapter plusieurs paramètres du pipeline.

Le lexique initial doit être redéfini en fonction du domaine étudié ; les sources documentaires doivent être sélectionnées selon leur pertinence thématique et leur accessibilité ; les seuils de filtrage, la fenêtre contextuelle et les critères de validation doivent être ajustés en fonction de la taille du corpus et des objectifs de recherche. Le pipeline décrit dans l'Algorithme 1 peut ainsi servir de cadre générique pour la constitution de corpus thématiques, depuis la collecte contrôlée des données jusqu'à la production de représentations vectorielles et de métadonnées réutilisables.

À titre indicatif, la phase de collecte et de nettoyage automatisé représente la part la plus importante du traitement computationnel, tandis que la validation experte constitue l'étape la plus coûteuse en temps humain. Dans le cas présenté, la faisabilité du protocole repose sur l'automatisation des étapes de repérage, d'extraction et de vectorisation, puis sur une validation humaine ciblée des résultats lexicaux. Cette répartition permet de limiter la charge de travail manuelle tout en maintenant un contrôle qualitatif sur les termes retenus.

5. Conclusion

En réponse aux besoins de construction de corpus destinés à l'identification de concepts émergents, la méthodologie proposée démontre la faisabilité d'une approche hybride, reproductible et évolutive pour la constitution de corpus multilingues thématiques.

Le dispositif associe extraction automatisée, structuration lexicale et annotation statistique, afin de rendre compte des régularités lexicales observées dans des textes issus de sources hétérogènes, tout en garantissant leur traçabilité.

Le jeu de données dérivé, intégralement vectorisé, constitue une ressource réutilisable et interopérable pour la recherche et l'apprentissage automatique. Il regroupe les blocs contextuels extraits autour des termes-clés identifiés, enrichis de métadonnées descriptives (secteur, langue, score TF-IDF, *n*-gramme) et d'un encodage vectoriel assurant la conformité éthique et juridique du partage. Ce format de diffusion préserve la valeur scientifique du corpus tout en éliminant les risques liés à la redistribution de contenus protégés.

Le *dataset* dérivé décrit dans cet article est disponible en ligne, dans le dépôt `malantin_dataset_release/` et via le DOI associé. Il correspond au dossier `dataset_release/`, qui regroupe les sous-ensembles `french/`, `english/` et `machine_learning_dataset/`. Les données diffusées comprennent uniquement les représentations vectorielles non directement réversibles et les métadonnées non textuelles ; les textes bruts issus des pages web et des rapports annuels ne sont pas redistribués.

L'intégration d'un apprentissage supervisé, combinée à une validation experte, a permis d'identifier et de confirmer de nouveaux termes liés à l'innovation non technologique, renforçant la robustesse du protocole

et sa portée interdisciplinaire. Cette approche offre une passerelle entre méthodes computationnelles et analyse linguistique, en rendant possible l'exploration systématique de dynamiques lexicales dans des champs conceptuels en évolution.

Au-delà du cas d'étude présenté, ce travail contribue à combler le manque de ressources adaptées aux besoins des sciences humaines et sociales, en facilitant leur réutilisation dans des cadres ouverts et collaboratifs.

Les perspectives incluent l'extension du protocole à d'autres domaines conceptuels, la normalisation des métadonnées selon les principes FAIR, et l'intégration progressive de ces corpus dans des infrastructures partagées telles que SDHSS (*Semantic Data for Humanities and Social Sciences*).

Ainsi, le projet illustre la complémentarité entre l'informatique et les sciences humaines, en croisant ingénierie linguistique et analyse des dynamiques d'innovation à partir de corpus textuels d'entreprises.

6. Remerciements

Ce travail s'inscrit dans le cadre du projet MALANTIN. Les auteures remercient Philippe Laredo, Tita Kyriacopoulou, Antoine Schoen et Patricia Laurens pour leurs échanges et leur appui dans la définition du cadre conceptuel. Elles remercient également le laboratoire LIASD (Université Paris 8) pour les moyens techniques mis à disposition, ainsi que leur collègue Florence Bannay (IRIT) pour ses remarques constructives sur la présentation méthodologique. Elles expriment enfin leur gratitude à Julien Muller (CNRS, comité de rédaction de Data & Corpus) pour ses conseils et suggestions éditoriales lors de la préparation de cet article, ainsi qu'aux stagiaires et partenaires du projet.

7. Déclaration d'intérêts ou d'absence d'intérêts conflictuels

Absence d'intérêts conflictuels.

8. Contributions

Anna Pappa : conceptualisation, méthodologie, ressources, développement, supervision, rédaction - ébauche du manuscrit, relectures et corrections.

Revekka Kyriakoglou : analyse formelle, ressources, développement, tests, validation, rédaction, relectures et corrections.

9. Références bibliographiques scientifiques

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209-226. https://doi.org/10.1007/s10579-009-9081-4

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising the WordNet Domains hierarchy: Semantics, coverage and balancing. Dans *Proceedings of the Workshop on Multilingual Linguistic Resources* (pp. 94-101). COLING. https://doi.org/10.3115/1706238.1706254

Beretta, F. (2024). Semantic data for humanities and social sciences (SDHSS): An ecosystem of CIDOC CRM extensions for research data production and reuse. Dans T. Riechert, H. Beyer, J. Blanke & E. Marx (dir.), *Professorale Karrieremuster Reloaded : Entwicklung einer wissenschaftlichen Methode zur Forschung auf online verfügbaren und verteilten Forschungsdatenbanken der Universitätsgeschichte* (pp. 73–101). HTWK Open-Access-Hochschulverlag. https://doi.org/10.33968/9783966270502-05

Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R., Simon, J., Swiezinski, L., & Zesch, T. (2013). Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics*, *28*(2), 23-59. https://doi.org/10.21248/jlcl.28.2013.175

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual WordNet. Dans *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers, pp. 1352-1362). Association for Computational Linguistics. https://aclanthology.org/P13-1133/

Conia, S., Barba, E., Martinez Lorenzo, A. C., Huguet Cabot, P.-L., Orlando, R., Procopio, L., & Navigli, R. (2024). MOSAICo: A multilingual open-text semantically annotated interlinked corpus. Dans *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers, pp. 7990-8004). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.442

Essien, I. A., Adebayo, A. O., Afuwape, A. A., Eboseremen, B. O., Oladega, F., & Soneye, O. M. (2023). The ethics of web scraping in research: Investigating the boundaries, legal implications, and societal acceptance of web scraping as a data collection method. *Journal of Frontiers in Multidisciplinary Research*, *4*(1), 529-538. https://doi.org/10.54660/.JFMR.2023.4.1.529-538

Gonzalez-Agirre, A., Castillo, M., & Rigau, G. (2012). A proposal for improving WordNet Domains. Dans *Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC'12*, (pp. 3457-3462). European Language Resources Association.

Han, L., Jones, G., & Smeaton, A. (2020). AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. Dans *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (pp. 44-57). Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.6/

Hinrichs, E., & Krauwer, S. (2014). The CLARIN research infrastructure: Resources and tools for eHumanities scholars. Dans *Proceedings of the Ninth International Conference on Language Resources and Evaluation – LREC'14*, (pp. 1525-1531). European Language Resources Association.

Ide, N., & Pustejovsky, J. (Eds.). (2017). *Handbook of linguistic annotation*. Springer. https://doi.org/10.1007/978-94-024-0881-2

Kazmali, A. S., & Sayar, A. (2025). Web scraping: Legal and ethical considerations in general and local context - A review. *Procedia Computer Science*, *259*, 1563-1572. https://doi.org/10.1016/j.procs.2025.04.111

Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010). A corpus factory for many languages. Dans *Proceedings of the Seventh International Conference on Language Resources and Evaluation – LREC'10*. European Language Resources Association. https://aclanthology.org/L10-1044/

Kyriakoglou, R., Pappa, A., He, J., Schoen, A., Laurens, P., Vartampetian, M., Laredo, P., & Kyriacopoulou, T. (2024). Développement automatique de lexiques pour les concepts émergents : Une exploration méthodologique. Dans A. Dister & D. Longrée (dir.), *JADT 2024 : Mots comptés, textes déchiffrés* (vol. 1, Cahiers du CENTAL, n° 17). Presses universitaires de Louvain. hal-04772221v1

Lui, M., & Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. Dans *Proceedings of the ACL 2012 System Demonstrations* (pp. 25–30). Association for Computational Linguistics. https://aclanthology.org/P12-3005/

Myers, D., & McGuffee, J. W. (2015). Choosing Scrapy. *Journal of Computing Sciences in Colleges*, *31*(1), 83-89.

Nakov, P. (2015). Web as a corpus: Going beyond the n-gram. Dans P. Braslavski, N. Karpov, M. Worring, Y. Volkovich, & D. I. Ignatov (dir.), *Information retrieval: 8th Russian Summer School, RuSSIR 2014*, Nizhniy Novgorod, Russia, August 18–22, 2014, revised selected papers (pp. 185-228). Springer. https://doi.org/10.1007/978-3-319-25485-2_5

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217-250. https://doi.org/10.1016/j.artint.2012.07.001

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S. ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pappa, A., & Kyriakoglou, R. (2025). *Données de réplication pour : MALANTIN* (Version V1) [Jeu de données]. Recherche Data Gouv. https://doi.org/10.57745/M9UHMX

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 31*(1), 4444-4451. https://doi.org/10.1609/aaai.v31i1.11164