

ARTICLES

Modélisation lexicale et intelligence artificielle : une ressource bilingue pour l'architecture contemporaine

Lexical modeling and artificial intelligence: A bilingual resource for contemporary architecture

Zaida Bartolomé-Díaz

Universidad de Las Palmas de Gran Canaria

zaida.bartolome@ulpgc.es

<https://orcid.org/0000-0002-7988-0586>

Abstract

This article presents a methodology for the design of a bilingual French-Spanish lexical resource dedicated to contemporary architecture, a field characterized by a constantly evolving terminology and a lack of specialized dictionaries. Our proposal is based on a model that integrates the entire lexicographic workflow – from corpus building to final encoding, including extraction, alignment, definition, and enrichment – while leveraging the contributions of artificial intelligence. Statistical and neural models are used to automate steps that are traditionally lengthy and demanding – such as term identification, semantic clustering, draft definition generation, or the detection of variants. However, this automation remains framed by systematic human validation, which is indispensable to ensure conceptual accuracy and terminological consistency. More broadly, our methodology provides a roadmap that can be applied to other specialized domains and to under-represented languages.

Keywords: specialized lexicography, bilingual resource, contemporary architecture, TEI-LMF modeling, artificial intelligence

RECONFIGURER LE TRAVAIL LEXICOGRAPHIQUE GRÂCE À L'IA

Portée par les technologies numériques et les progrès du traitement automatique du langage, la lexicographie s'est profondément transformée au cours des dernières années. Ce mouvement atteint aujourd'hui une nouvelle étape, marquée par l'essor de l'intelligence artificielle. Grâce à l'IA, et plus particulièrement à l'apprentissage profond et aux modèles de type *transformers*, il est aujourd'hui possible de traiter des tâches clés comme l'extraction de définitions (Periti *et al.*, 2024), l'alignement bilingue et l'enrichissement sémantique des unités lexicales (Cook, 2020), ou encore la génération de contenus lexicographiques (Lew, 2023, 2024). Dans cette perspective, De Schryver (2023) rappelle que l'IA ne vise pas à remplacer le travail lexicographique, mais à l'amplifier en facilitant une analyse plus rapide et plus systématique de données massives, ce qui a profondément reconfiguré le rôle du lexicographe, désormais moins compilateur de définitions que véritable analyste et médiateur de données linguistiques (Grabowski, 2023).

Bien que plusieurs travaux récents aient exploré l'usage de l'intelligence artificielle pour générer des dictionnaires ou des composantes lexicographiques – par exemple le *Spanish-BFF* (Ortega-Martín *et al.*, 2023), sa version améliorée avec GPT-4 (*Spanish-BFF-2.0*, Ortega-Martín *et al.*, 2024), la génération de phrases exemples bilingues assistées par des modèles de grande taille (Merx, Vylomova & Kurniawan, 2024), ou encore l'expérimentation de Jakubíček et Rundell (2023) avec un dictionnaire de 99 entrées produit par ChatGPT – il n'existe pas, à notre connaissance, de feuille de route méthodologique couvrant l'ensemble du processus : de l'extraction et la sélection des unités lexicales à leur compilation et encodage final selon des standards interopérables.

De ce fait, cet article propose d'aborder cette question en explorant les principes de conception d'une ressource lexicale bilingue français-espagnol dédiée à l'architecture contemporaine. L'objectif est de proposer un modèle méthodologique qui articule l'ensemble du travail lexicographique – de l'identification des unités à leur structuration – avec les standards d'encodage et l'apport de l'intelligence artificielle, afin de garantir la rigueur linguistique, la cohérence conceptuelle et l'interopérabilité technique.

1. L'INTELLIGENCE ARTIFICIELLE COMME LEVIER MÉTHODOLOGIQUE

Loin d'être un simple outil d'assistance, l'IA est envisagée aujourd'hui dans le travail lexicographique comme un véritable levier méthodologique, car elle permet d'automatiser des opérations complexes, de traiter des volumes massifs de données textuelles et de dégager des régularités difficilement observables par une analyse manuelle.

Ainsi, le *Spanish Built Factual Freectionary* (Ortega-Martín *et al.*, 2023) a montré que GPT-3 pouvait générer automatiquement des entrées initiales, tandis que sa version améliorée, *Spanish-BFF-2* (Ortega-Martín *et al.*, 2024), exploitant GPT-4-turbo, a démontré une plus grande richesse lexicale et une réduction des erreurs typiques comme les hallucinations ou les définitions trop vagues. Dans une autre perspective, *Dict-BERT* (Yu *et al.*, 2022) illustre comment les définitions issues de dictionnaires peuvent enrichir la représentation des mots rares dans les modèles de langage.

D'autres travaux proposent des créations automatiques de dictionnaires thématiques à partir du Web, comme celui de Toprak et Turan (2025), qui présentent un modèle qui exploite des similitudes et ressources en ligne pour générer des ensembles lexicographiques.

De leur côté, Li et Tarp (2024) ont expérimenté l'usage de chatbots (ChatGPT, Ernie Bot) pour fournir des explications lexicographiques à des apprenants chinois d'anglais, confirmant le potentiel de ces outils à générer des contenus pédagogiques utiles, mais aussi la nécessité d'une validation experte. Lew (2024) rappelle dans la même ligne que les LLMs¹ ne sauraient se substituer aux lexicographes : ils doivent rester des aides puissantes, mais guidées par un jugement humain garantissant la nuance, l'adéquation contextuelle et l'éthique. Enfin, Fung *et al.* (2024) montrent, dans le cadre de l'annotation semi-automatique de documents, que les approches hybrides « human-in-the-loop » sont les plus fiables car l'humain intervient pour corriger les erreurs générées automatiquement.

Ces contributions récentes convergent vers une même conclusion : l'IA ouvre des perspectives inédites pour accélérer et enrichir le travail lexicographique, mais sa pleine valeur ne s'exprime que lorsqu'elle est intégrée dans un processus collaboratif où l'expertise humaine demeure garante de la qualité scientifique et de la pertinence conceptuelle des ressources.

2. LA LEXICOGRAPHIE SPÉCIALISÉE DE L'ARCHITECTURE SOUS LA LOUPE

Dans ce paysage en mutation, la lexicographie spécialisée occupe une place essentielle : elle vise à documenter et à systématiser la terminologie propre à des domaines techniques et professionnels, en tenant compte des besoins communicationnels concrets des usagers (Bergenholtz & Tarp, 1995). La constitution de ressources bilingues s'avère particulièrement complexe, car elle exige non seulement une maîtrise linguistique approfondie, mais aussi une connaissance du domaine thématique et des outils numériques adaptés (L'Homme, 2019).

¹ *Large Language Model* ou grands modèles de langage.

Le domaine de l'architecture contemporaine illustre parfaitement cette complexité. Sa terminologie évolue rapidement sous l'effet des innovations technologiques, de l'introduction de nouveaux matériaux, mais aussi de l'émergence de concepts liés au développement durable et à l'écoconstruction (Bartolomé-Díaz & Trujillo-González, 2023). Ces transformations ne concernent pas seulement les objets matériels, mais traduisent également des changements conceptuels majeurs qui reconfigurent la manière de penser l'espace bâti et urbain (Durand, 2023).

Or, une lacune significative subsiste concernant les ressources bilingues français-espagnol spécialisées dans ce champ (Bartolomé-Díaz & Trujillo-González, 2023). Les dictionnaires généraux ou les ressources et glossaires techniques disponibles ne couvrent ni l'étendue ni la dynamique de la terminologie architecturale actuelle. Cette carence entrave non seulement la recherche et l'enseignement, mais aussi la pratique professionnelle, où la précision terminologique est déterminante pour la conception, la construction et la documentation des projets architecturaux. En effet, l'effort de répertorier et d'organiser les unités lexicales de l'architecture ne constitue pas seulement un enjeu académique : il répond également à une nécessité concrète pour les professionnels et chercheurs confrontés à des échanges multilingues. Dans cette perspective, Becchi *et al.* (2008) soulignent que la constitution de vocabulaires, glossaires et dictionnaires dans le domaine de l'architecture représente l'un des défis majeurs pour les années à venir.

De plus, les ressources existantes, lorsqu'elles existent, ne sont pas toujours conçues selon des standards internationaux d'encodage ouverts et d'interopérabilité. L'absence d'une modélisation rigoureuse limite leur exploitation dans des infrastructures numériques ou leur intégration dans des flux de travail avancés, tels que ceux impliquant l'intelligence artificielle.

Ainsi, la diversité et l'évolution constante de la terminologie de l'architecture, la rareté de ressources bilingues dans le domaine et la nécessité de documenter les nouveaux concepts liés à l'innovation en font un secteur particulièrement représentatif des enjeux actuels de la lexicographie spécialisée. De ce fait, le champ de l'architecture contemporaine nous est apparu comme un terrain privilégié pour concevoir et mettre à l'épreuve une méthodologie de constitution d'une ressource lexicographique assistée par l'IA.

3. PRINCIPES DE CONCEPTION DE NOTRE RESSOURCE LEXICALE

De notre point de vue, la conception d'une ressource lexicale spécialisée doit être envisagée dès l'origine comme un projet scientifique durable, interopérable et réutilisable. La simple compilation d'entrées ne suffit pas, nous devons garantir que les données pourront circuler entre différentes infrastructures et être exploitées sur le

long terme. Comme le souligne Romary (2011), la stabilisation des savoirs dans les humanités passe par leur inscription dans des standards partagés, condition essentielle pour assurer la pérennité et la compatibilité des ressources.

Dans cette perspective, notre travail se fonde sur trois principes méthodologiques :

- Fondation empirique : les données lexicales doivent être ancrées dans l’usage réel, ce qui implique de constituer des corpus spécialisés représentatifs et de documenter les unités dans leurs contextes authentiques.
- Approche bilingue intégrée : la ressource est conçue dès l’origine comme un outil bilingue (français-espagnol), ce qui exige non seulement l’identification des termes, mais aussi leur mise en relation systématique et leur harmonisation conceptuelle.
- Normalisation et interopérabilité : la modélisation des données s’appuie sur des standards internationaux, garantissant leur compatibilité avec les infrastructures linguistiques existantes et leur potentiel d’évolution.

4. ÉTAPES D’UN FLUX HYBRIDE : UNE MÉTHODOLOGIE ASSISTÉE PAR L’IA

Dans la continuité des principes précédemment exposés, notre démarche s’appuie sur un flux méthodologique hybride, où l’intelligence artificielle intervient à différents niveaux du processus lexicographique. Cette hybridation ne renvoie pas seulement à l’articulation entre outils numériques et expertise humaine, mais également à la complémentarité des types de modèles mobilisés selon la nature des opérations à accomplir.

Ainsi, les modèles statistiques, tels que ceux intégrés dans SketchEngine (Kilgarriff *et al.*, 2004), reposent sur des mesures de fréquence, de spécificité et de cooccurrence permettant d’extraire et de hiérarchiser les candidats lexicaux. Les modèles distributionnels (*word embeddings*) apprennent, pour leur part, à représenter les mots sous forme de vecteurs dans un espace continu où la proximité géométrique traduit la proximité sémantique. Les modèles multilingues, tels que MUSE (Artetxe *et al.*, 2018), prolongent cette logique en projetant des représentations issues de plusieurs langues dans un espace partagé, permettant d’établir des correspondances automatiques entre lexiques. Les modèles neuronaux contextuels – comme CamemBERT (Martin *et al.*, 2020) ou BETO (Cañete *et al.*, 2020) – offrent des représentations vectorielles sensibles au contexte, utiles pour détecter des proximités sémantiques fines ou des variations d’usage entre corpus.

Les grands modèles de langage (LLMs), tels que GPT-4, combinent apprentissage massif et raisonnement textuel, et peuvent être exploités pour la génération contrôlée de définitions ou de reformulations bilingues. Enfin, leurs applications conversationnelles, comme ChatGPT, ouvrent de nouvelles perspectives en matière d’interaction lexicographique et d’accompagnement pédagogique.

Sur cette base, notre travail lexicographique se déploie en plusieurs étapes, chacune correspondant à une opération qui, dans une approche traditionnelle, exigerait un investissement considérable en temps et en ressources, mais qui a pu être rationalisée grâce à l’usage combiné de ces différents modèles :

- Extraction terminologique augmentée à l’aide de modèles statistiques et distributionnels.
- Alignement bilingue grâce aux modèles multilingues.
- Classification sémantique avec des modèles contextuels.
- Génération intelligente de définitions et d’exemples à partir des LLMs.
- Enrichissement sémantique guidé par l’IA via des modèles distributionnels et contextuels.
- Automatisation de la modélisation assistée par des scripts d’IA générative pour la structuration et la validation des données.

Voyons plus en détails chacune de ces étapes.

4.1. EXTRACTION TERMINOLOGIQUE AUGMENTÉE : DE SKETCHENGINE AUX MODÈLES NEURONAUX

Le point de départ de notre travail a été la constitution de deux corpus parallèles (français et espagnol) avec SketchEngine² (Kilgarriff *et al.*, 2004), couvrant environ vingt-cinq ans de textes spécialisés en architecture contemporaine.

L’extraction initiale des candidats termes a été réalisée directement avec cet outil, lequel combine des méthodes statistiques de fréquence et de spécificité (*keywords*, *term lists*) ainsi que des analyses de cooccurrences (*word sketches*). Ces fonctions permettent d’isoler les unités qui se distinguent de la langue générale par leur forte représentativité dans le domaine spécialisé.

Dans notre corpus, cela a conduit à la détection de formes comme *béton auto-citrifiant*, *ossature bois* ou *plancher champignon* en français, et *hormigón autocompactante* ou *estructura de madera* en espagnol. Ces unités, simples ou polylexicales, ont été extraites à l’aide des méthodes statistiques et morphosyntaxiques mises en œuvre dans SketchEngine (n-grammes, patrons nominaux), et ont constitué la base brute de la ressource.

Cependant, cette extraction ne suffisait pas à garantir la pertinence terminologique des unités identifiées. Certaines formes, comme *structure*, *projet* ou *système*,

² Le choix de Sketch Engine pour la constitution des corpus et l’extraction initiale des unités lexicales repose sur la nécessité de garantir la fiabilité empirique et la transparence méthodologique des données. Contrairement aux approches fondées exclusivement sur l’intelligence artificielle, qui génèrent ou sélectionnent des textes selon des critères souvent opaques, Sketch Engine permet de travailler sur des corpus attestés, équilibrés et documentés, dont les métadonnées (source, date, domaine, type de texte) sont explicitement identifiables.

bien que fréquentes et statistiquement spécifiques dans nos textes techniques, appartiennent déjà au vocabulaire général du discours scientifique et ne possèdent plus de valeur discriminante pour le domaine architectural.

De ce fait, pour affiner la sélection, nous avons donc cherché à mesurer la spécificité contextuelle de chaque candidat à terme à l'aide de modèles distributionnels (*word embeddings*) capables de représenter la proximité sémantique entre les mots en fonction de leurs cooccurrences.

Chaque unité a ainsi été représentée par deux profils distributionnels distincts, obtenus à l'aide de modèles fastText (Bojanowski *et al.*, 2017) entraînés séparément sur nos corpus spécialisés français et espagnol et sur des corpus de langue générale (*frWaC* pour le français, *esTenTen* pour l'espagnol) afin d'évaluer la divergence entre les contextes d'usage dans les deux registres.

Pour chaque candidat, nous avons extrait, dans chacun des corpus (spécialisé et général), la liste des k voisins les plus proches dans l'espace vectoriel, constituant ce que l'on appelle son profil distributionnel, entendu comme l'ensemble des mots avec lesquels il tend à apparaître dans un contexte donné. Ces profils distributionnels ont ensuite été comparés afin d'évaluer la divergence entre les contextes d'usage dans les deux registres³.

Cette comparaison repose sur la mesure de similarité cosinus, qui quantifie l'angle entre les vecteurs représentant un même mot dans deux espaces différents : un cosinus proche de 1 indique une grande proximité entre les deux profils – typique des unités à usage transversal ou générique –, tandis qu'un cosinus faible signale une divergence contextuelle forte, indicatrice d'une spécialisation thématique accrue.

Ainsi, par exemple, le terme-candidat *structure* a un cosinus entre les deux corpus élevé ($\approx 0,89$), ce qui montre que ses cooccurrences – *système, organisation, modèle, plan...* – demeurent similaires d'un contexte à l'autre. Cette stabilité traduit un usage polysémique et transversal : le mot conserve une fonction métalinguistique dans de nombreux domaines techniques, sans renvoyer spécifiquement à un concept architectural précis. Il a donc été exclu de la liste finale.

En revanche, le terme-candidat *plancher champignon* présente un profil radicalement différent. Son cosinus entre le corpus général et le corpus spécialisé est faible ($\approx 0,35$), indiquant une forte différenciation contextuelle. Dans notre corpus d'architecture, ses voisins – *dalle, poteau, chapiteau, béton armé, porteur* – décrivent des éléments structurels concrets ; dans la langue générale, en revanche, le terme est presque absent. Ce décalage confirme sa valeur terminologique spécifique et justifie son maintien dans la ressource.

Cette différenciation nous a permis de conserver uniquement les unités dont la spécificité contextuelle confirmait l'appartenance au champ de l'architecture, tout en

³ La comparaison ne porte donc pas sur les vecteurs bruts eux-mêmes, qui ne sont pas directement comparables d'un espace vectoriel à l'autre, mais sur les profils distributionnels ainsi construits.

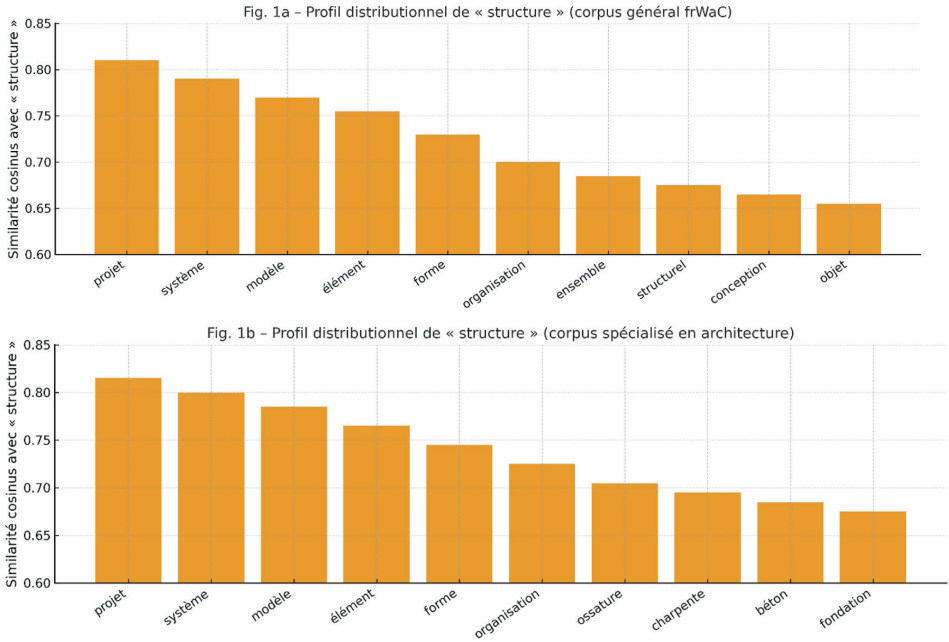


Figure 1. Comparaison des profils distributionnels du mot *structure*

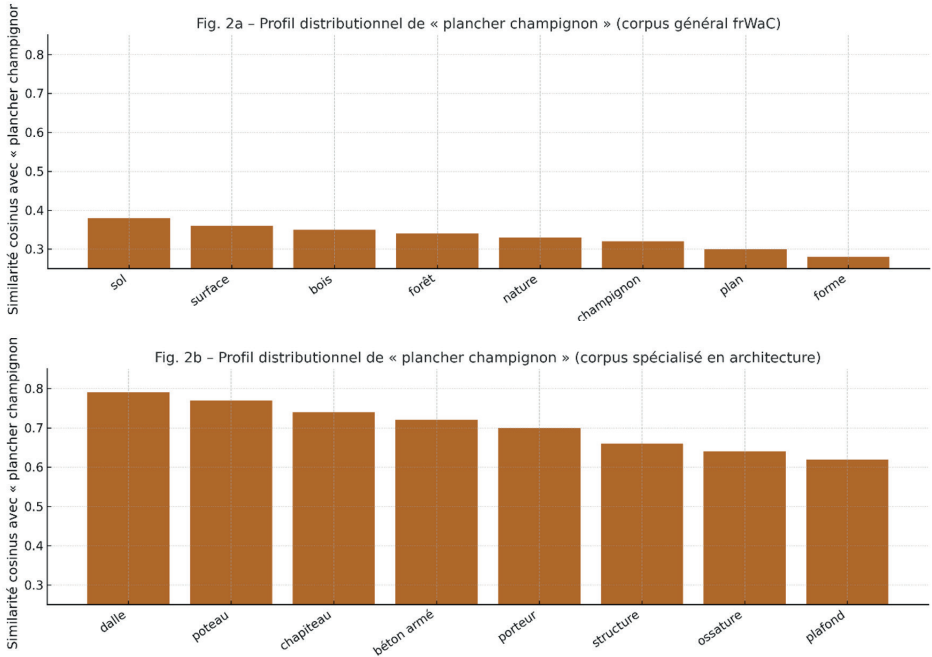


Figure 2. Comparaison des profils distributionnels du mot *structure* (corpus spécialisé vs. corpus général)

écartant celles dont le comportement distributionnel témoignait d'un usage trop général ou interdisciplinaire.

Ainsi, l'IA n'a pas remplacé SketchEngine, mais a joué le rôle de filtre de spécialisation, réduisant la charge de validation manuelle et augmentant la précision de la liste terminologique finale.

4.2. ALIGNEMENT BILINGUE AVEC DES MODÈLES MULTILINGUES

À partir des listes monolingues des candidats à termes issues de SketchEngine et du filtrage distributionnel, l'étape suivante a consisté à établir des correspondances entre les termes français et espagnols. Cette opération a été réalisée grâce à des modèles multilingues capables de projeter les représentations vectorielles de mots appartenant à différentes langues dans un espace sémantique partagé.

Concrètement, nous avons utilisé des méthodes de transfert non supervisé (Artetxe *et al.*, 2018) mises en œuvre dans des bibliothèques comme MUSE (*Multilingual Unsupervised and Supervised Embeddings*), qui alignent automatiquement des vecteurs issus de deux langues différentes. Cette approche repose sur une transformation linéaire orthogonale permettant de rapprocher les deux espaces vectoriels et de placer les unités lexicales de sens équivalent à proximité dans une même dimension multilingue.

Dans ce cadre :

- Chaque terme français et espagnol extrait a été représenté par son vecteur distributionnel (établis séparément à l'étape précédente).
- Les modèles de transfert ont ensuite appliqué une transformation linéaire qui rapproche les deux espaces vectoriels dans une même dimension multilingue.
- L'appariement s'effectue ensuite par mesure de proximité cosinus : le terme espagnol dont le vecteur est le plus proche d'un terme français est proposé comme équivalent.

Par exemple, le modèle a aligné avec succès *ossature bois* ↔ *estructura de madera* et *béton précontraint* ↔ *hormigón pretensado*. Ces correspondances automatiques ont ensuite été vérifiées et validées par expertise humaine, avant son intégration dans la ressource.

4.3. CLASSIFICATION SÉMANTIQUE AVEC MODÈLES CONTEXTUELS

Après l'alignement bilingue, l'étape suivante a consisté à organiser les termes en réseaux sémantiques afin de rendre explicites leurs relations conceptuelles sous-jacents dans le corpus. Pour cela, nous avons mobilisé des modèles neuronaux contextuels de type BERT (Devlin *et al.*, 2019), en exploitant les ressources disponibles

CamemBERT (Martin *et al.*, 2020) pour le français et BETO pour l’espagnol (Cañete *et al.*, 2020), ainsi que des approches distributionnelles (Lenci, 2018).

Contrairement aux *embeddings* statiques, ces modèles génèrent pour chaque occurrence d’un mot une représentation vectorielle dépendante du contexte. Autrement dit, un même terme n’a pas une seule représentation fixe, mais plusieurs, qui varient selon son environnement syntaxique et sémantique. Cette propriété permet de capturer les nuances d’usage et les relations de proximité conceptuelle entre termes appartenant à un même domaine.

Concrètement, chaque phrase contenant une unité lexicale, simple ou polylexicale, a été tokenisée selon le schéma propre au modèle, puis encodée afin de produire des représentations vectorielles contextualisées⁴.

L’ensemble de ces représentations contextuelles a permis de projeter les unités lexicales dans un espace vectoriel fondé sur leurs occurrences réelles dans le corpus. La comparaison de ces représentations a ensuite servi à identifier des proximités sémantiques entre unités partageant des environnements d’usage similaires et à proposer des regroupements conceptuels. Ainsi, *toiture végétalisée*, *toit vert* et *toiture extensive* se sont rapprochés pour former un cluster, suggérant une hiérarchie conceptuelle confirmée ensuite par expertise humaine. De même, les modèles ont mis en évidence la proximité entre *béton armé* et *béton précontraint*, reflétant leur appartenance à une même famille sémantique.

4.4. GÉNÉRATION INTELLIGENTE DE DÉFINITIONS ET D’EXEMPLES

Pour enrichir les entrées lexicales, nous avons eu recours à des grands modèles de langage (LLMs), en particulier GPT-4 (OpenAI, 2023), un modèle de très grande capacité entraîné sur des corpus massifs et conçu pour produire du texte cohérent à partir d’instructions en langage naturel. L’objectif n’était pas de produire des définitions « clé en main », mais de générer des ébauches à partir des contextes les plus fréquents.

Concrètement, nous avons expérimenté son application conversationnelle, ChatGPT⁵, adaptée à notre domaine grâce à la conception de prompts guidés construits à partir de contextes authentiques extraits de notre corpus spécialisé avec *SketchEngine*. Cette adaptation consiste à orienter le modèle vers un usage spécialisé en l’exposant à des exemples issus de nos corpus d’architecture, afin de réduire les généralisations trop larges et d’obtenir des formulations en adéquation avec le domaine.

⁴ Pour les unités simples, le vecteur associé au token correspondant a été extrait ; pour les unités polylexicales, la représentation a été établie à partir de la séquence de tokens correspondant à l’unité dans la phrase.

⁵ Pour la génération et la reformulation, nous avons eu recours à la version gratuite de ChatGPT-4.

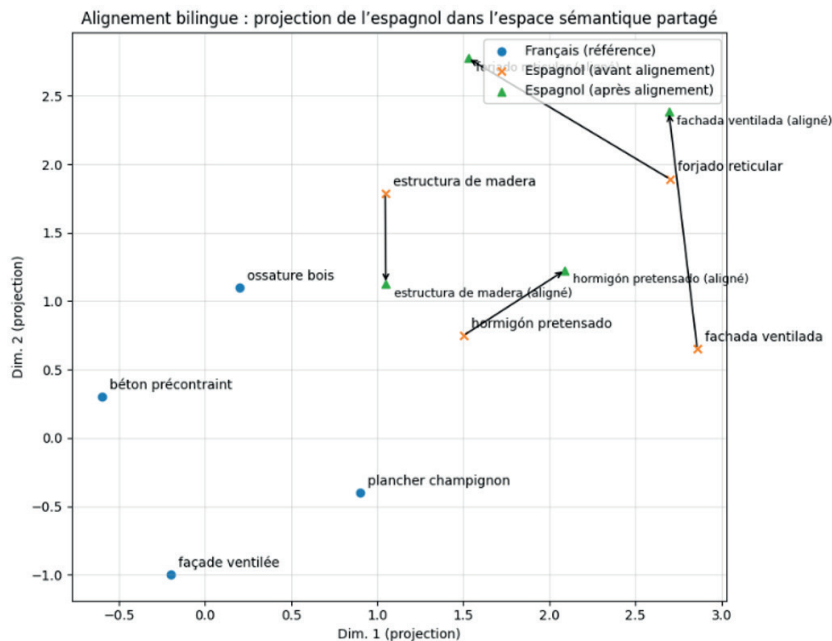


Figure 3. Alignement bilingue : projection des termes français et espagnols dans l'espace sémantique partagé

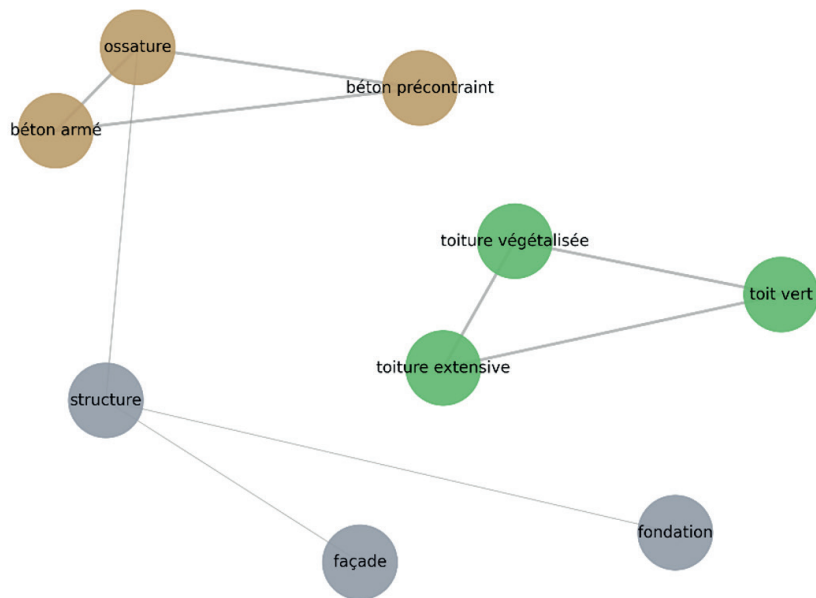


Figure 4. Regroupements sémantiques (CamemBERT/BETO)

Pour chaque terme, plusieurs contextes d'occurrence ont ainsi été sélectionnés et soumis au modèle afin de générer :

- une définition provisoire synthétisant les régularités d'usage ;
- des exemples reformulés à partir des cooccurrences les plus fréquentes.

L'IA a ainsi été mobilisée pour synthétiser et reformuler l'information issue des contextes d'usage, proposant une synthèse lexicographique que nous avons ensuite validée et réécrite.

Pour rendre ce processus reproductible, nous avons défini des *prompts* guidés comprenant :

1. le terme à décrire,
2. un échantillon de contextes authentiques issus de notre corpus spécialisé,
3. une instruction explicite du type « *Résume les informations communes pour proposer une définition courte et neutre adaptée à un dictionnaire spécialisé d'architecture* ».

Le prompt a été progressivement affiné par itérations successives, en ajustant sa formulation, la longueur des contextes et le degré de contrainte sémantique. Nous avons constaté qu'un prompt trop général produisait des définitions trop encyclopédiques, tandis qu'un prompt plus ciblé sur le champ technique générait des descriptions plus précises mais parfois elliptiques.

Par exemple, pour *béton armé*, nous avons d'abord utilisé la consigne générale – « *Résume les informations communes ...* » –, ce qui produisait des reformulations trop encyclopédiques, souvent accompagnées d'éléments historiques ou métaphoriques.

En ajoutant ensuite une contrainte de forme (« *Limite ta réponse à une seule phrase, en adoptant un ton neutre et descriptif, sans information contextuelle externe* »), nous avons obtenu des définitions plus concises, mais encore trop vagues techniquement.

Enfin, en précisant le champ d'application (« *Formule une définition courte, adaptée à un dictionnaire d'architecture, en mettant l'accent sur la fonction structurelle du matériau* »), le modèle a produit une synthèse terminologique correcte, où l'idée de renforcement par armature métallique apparaissait clairement.

À partir de ce prompt stabilisé et les contextes d'usage extraits avec *SketchEngine*, ChatGPT a condensé l'information en mettant en évidence que la fonction principale du matériau résidait dans le renforcement de la résistance mécanique par l'insertion d'armatures et nous avons pu obtenir une définition⁶ :

« *Béton renforcé par une armature métallique afin d'améliorer sa résistance mécanique.* »

Ce processus montre comment les grands modèles de langage peuvent devenir de véritables outils d'assistance méthodologique, capables de condenser l'information

⁶ L'expérimentation de ces méthodes de génération assistée de définitions est encore en cours. Leur évaluation systématique et leur intégration dans la ressource feront l'objet d'une publication distincte consacrée spécifiquement à ce volet.

issue de multiples contextes, d'en faciliter la reformulation et de fournir une base directement exploitable par le chercheur.

4.5. ENRICHISSEMENT SÉMANTIQUE GUIDÉ PAR L'IA

Les étapes précédentes ont permis d'extraire, d'aligner et de classer les unités lexicales selon leurs proximités contextuelles. Toutefois, les relations établies entre termes demeuraient parfois trop locales et dépendantes des cooccurrences observées dans les corpus. Certains concepts proches ne se reliaient pas entre eux faute de contextes communs suffisants, ce qui pouvait entraîner des lacunes dans le réseau lexical.

Afin de dépasser ces limites et de conférer une plus grande cohérence sémantique du réseau lexical de notre ressource, nous avons eu recours à des modèles multilingues enrichis, tels que SensEmBERT (Scarlini *et al.*, 2020), capables d'intégrer des informations lexicales préexistantes issues de ressources de référence.

Contrairement aux *embeddings* classiques, ces modèles combinent l'information contextuelle issue des corpus et les liens déjà encodés dans des bases lexicales comme WordNet (Miller, 1995) ou BabelNet (Navigli & Ponzetto, 2012), ce qui améliore la capacité à identifier des proximités sémantiques pertinentes. Ainsi, chaque terme de notre liste validée a été projeté dans cet espace vectoriel enrichi, puis comparé aux autres entrées de la ressource.

Le seuil de similarité a été défini empiriquement, en suivant le principe de comparaison vectorielle fondé sur la similarité cosinus décrit par Scarlini *et al.* (2020). Dans notre cas, les paires de termes présentant un score supérieur à 0,70 ont été considérées comme suffisamment proches pour justifier une proposition de relation lexicale (synonymie, hyponymie ou hyperonymie).

Ces suggestions se sont traduites par des propositions comme :

- ajouter *toit vert* comme synonyme de *toiture végétalisée*,
- rapprocher *ossature* et *estructura portante* par une relation d'équivalence,
- relier *plancher champignon* à *dalle* comme hyponyme.

Les nouveaux termes ainsi identifiés ont été intégrées après validation humaine, renforçant la cohérence sémantique de la ressource.

4.6. MODÉLISATION TEI-LMF AUTOMATISÉE PAR SCRIPTS GÉNÉRÉS PAR L'IA

Une fois les données lexicales produites et validées, il était indispensable de les formaliser dans un cadre normalisé afin d'en garantir la pérennité et l'interopérabilité.

Dans cette dernière étape, nous avons mobilisé des outils d'intelligence artificielle pour accompagner et automatiser partiellement le processus de structuration,

en facilitant le passage des unités validées vers un format conforme aux standards lexicographiques internationaux.

Pour assurer cette formalisation, notre ressource repose sur le langage XML, qui permet de structurer hiérarchiquement les composantes d’une entrée lexicale (terme, définition, équivalent, contexte d’usage, informations grammaticales, *etc.*). Cette approche offre la flexibilité nécessaire pour adapter la base à notre domaine d’étude et l’enrichir au fil de l’évolution terminologique.

S’inscrivant dans cette logique, Romary (2010, 2013, 2019) a montré que la TEI pouvait constituer une sérialisation complémentaire de LMF, couvrant les besoins spécifiques de l’encodage lexicographique et assurant une passerelle entre différents projets. Cette convergence, entérinée dans la restructuration du standard (Romary *et al.*, 2019), garantit la pérennité, la réutilisabilité et l’interopérabilité de la ressource.

C’est à ce stade que l’IA intervient, en contribuant à la formalisation automatisée de chaque entrée, conçue comme une unité autonome et interconnectée. Chaque fiche intègre les formes française et espagnole, une définition contextualisée, des exemples d’usage et des relations sémantiques (synonymie, hyperonymie, hyponymie), générées ou préremplies à l’aide de scripts Python et de *prompts* d’IA générative.

Les principales opérations d’automatisation mises en œuvre dans ce processus sont présentées ci-dessous :

- Pré-structuration des entrées : à partir des listes de termes validés, l’IA a généré automatiquement des gabarits XML incluant les éléments de base (<entry>, <form>, <sense>), ce qui a réduit le travail manuel répétitif et limité les risques d’incohérences.

Nous avons utilisé un script Python qui lisait les listes validées de termes au format CSV et produisait des fichiers XML de base :

```
python

for term in validated_terms:
    entry = f"""
    <entry xml:id="{term['id']}">
        <form type="lemma">
            <orth>{term['lemma']}</orth>
        </form>
        <sense>
            <def></def>
        </sense>
    </entry>"""
    write_to_xml(entry)
```

Script 1. Génération des gabarits XML à partir des listes de termes validés

Pour garantir la conformité du schéma, les premières versions des gabarits ont été générées via un prompt sur ChatGPT du type :

« Crée une entrée TEI-LMF vide comprenant <entry>, <form>, <sense>, <def>, prête à être complétée pour le terme X. »

Ce double dispositif (Python pour la boucle, IA pour la validation de structure) a permis d'obtenir un fichier de départ homogène et conforme au schéma TEI-LMF.

- Insertion semi-automatique des attributs : Les informations grammaticales ont été injectées automatiquement dans la balise <gramGrp> à partir d'annotations morphologiques extraites des corpus (*POS*, *genre*, *nombre*), réduisant ainsi une étape habituellement longue et répétitive à une simple vérification par le lexicographe.

Un script Python exploitant la bibliothèque *spaCy*⁷ et les sorties du modèle *CamemBERT* a pré-rempli les balises :

```
python

if token.pos_ == "NOUN":
    pos = "<pos>nom</pos>"
    if token.morph.get("Gender") == ["Masc"]:
        gen = "<gen>m</gen>"
```

Script 2. Remplissage automatique des balises <gramGrp> à partir des annotations de corpus

Ces fragments ont ensuite été intégrés automatiquement dans chaque <gramGrp> correspondant.

Dans certains cas ambigus, nous avons demandé à ChatGPT de compléter les valeurs manquantes à partir d'exemples du corpus, avec un prompt de vérification :

« Quelle est la catégorie grammaticale et le genre du mot suivant, dans un contexte d'architecture ? Réponds uniquement par les balises TEI correspondantes. »

- Alignement formel des équivalents bilingues : les correspondances obtenues grâce aux modèles multilingues (cf. 5.2) ont été exportées dans un fichier intermédiaire (CSV/JSON). Un script Python a converti ces données en fragments XML (TEI-LMF) de type <form type=«translation»>, selon le modèle suivant :

⁷ *spaCy* est une bibliothèque open-source de traitement automatique du langage en Python, optimisée pour certaines tâches du NLP (tokenisation, étiquetage morphosyntaxique – POS tagging –, lemmatisation, parsing, etc.).

```
python

for pair in alignments:
    print(f'<form type="translation" xml:lang="es">{pair["spanish"]}</form>')
```

Script 3. Conversion des paires alignées en fragments XML de type <form type=«translation»>

- Mécanisme de détection de doublons et de variantes internes : Une difficulté récurrente dans la constitution de bases lexicales est la redondance des unités : un terme peut apparaître sous plusieurs variantes orthographiques (*hormigón autocompactante* / *hormigón autocompactado*), morphologiques (*ossature bois* / *ossatures bois*), ou comme synonyme quasi-équivalent (*toit vert* / *toiture végétalisée*).

Pour automatiser cette vérification, nous avons exploité à nouveau les *embeddings* contextuels produits par CamemBERT (Martin *et al.*, 2020) pour le français et BETO (Cañete *et al.*, 2020) pour l'espagnol. Concrètement :

1. Toutes les entrées déjà encodées en TEI-LMF ont été projetées dans un espace vectoriel.
2. Lorsqu'un nouveau candidat à terme issu de l'extraction devait être ajouté, il était comparé automatiquement à l'ensemble des vecteurs existants au moyen d'une mesure de similarité cosinus.

```
python

similarity = cosine_similarity(vec_new, vec_existing)
if similarity > 0.85:
    alert_duplicate(term_new, term_existing)
```

Script 4. Détection automatique de doublons et de variantes internes (seuil de similarité fixé à 0,85).

Si le score dépassait un seuil prédéfini (0,85⁸), le système générait une alerte de proximité.

Deux types de situations pouvaient être alors identifiés

- Synonymie / quasi-synonymie, comme *toit vert* et *toiture végétalisée*.
- Variante orthographiques ou morphologiques, comme *hormigón autocompactante* et *hormigón autocompactado*.

Dans ces situations, une balise TEI <xr> était générée automatiquement, pointant vers l'entrée concernée grâce à son identifiant @target.

⁸ Nous avons fixé le seuil de similarité cosinus à 0,85, en nous appuyant sur les pratiques courantes de la littérature en vectorisation lexicale (Mikolov *et al.*, 2013 ; Camacho-Collados & Pilehvar, 2018), ce qui a permis de limiter les faux positifs tout en capturant les proximités pertinentes.

```
<xr type="variant" target="#hormigon_autocompactado"/>
```

Script 5. Comparaison vectorielle des entrées encodées et génération de liens <xr>

Ce fonctionnement équivalait à créer un « chemin de contrôle » dynamique : au lieu de parcourir manuellement la liste des termes existants pour vérifier les redondances, l'IA interrogeait directement la base en construction et renvoyait des alertes de proximité. Le rôle du lexicographe était de valider ou corriger la relation suggérée, ce qui a considérablement réduit le temps de traitement.

4.6.1. RÉSULTAT ET STRUCTURE D'UNE ENTRÉE GÉNÉRÉE AVEC ASSISTANCE DE L'IA

La mise en œuvre de ce flux automatisé aboutit à la génération d'entrées XML complètes et normalisées. L'exemple ci-dessous montre la structure XML générée automatiquement pour l'entrée *plancher champignon*, où l'IA a pré-rempli les champs morphosyntaxiques, les équivalents bilingues et les relations sémantiques selon le modèle TEI-LMF validé.

```
<entry xml:id="plancher_champignon" xmlns="http://www.tei-c.org/ns/1.0">
  <form type="lemma">
    <orth xml:lang="fr">plancher champignon</orth>
    <gramGrp>
      <pos>nom</pos>
      <gen>masculin</gen>
    </gramGrp>
  </form>

  <form type="translation">
    <orth xml:lang="es">forjado reticular</orth>
  </form>

  <sense xml:id="s1">
    <def>
      Plancher en béton armé sans poutres apparentes,
      où la dalle repose directement sur des colonnes à chapiteau évasé.
    </def>

    <cit type="example">
      <quote>
        Le plancher champignon permet une plus grande flexibilité
        dans l'aménagement intérieur des bâtiments.
      </quote>
    </cit>

    <re type="hyperonymie" ref="#plancher"/>
    <xr type="relatedEntry" target="#dalle-n"/>
  </sense>
</entry>
```

Script 6. Encodage TEI-LMF de l'entrée *plancher champignon*

5. WORKFLOW FINAL DE CONSTITUTION DE LA RESSOURCE LEXICALE

Le schéma ci-dessous présente de manière synthétique, l'ensemble des étapes méthodologiques décrites dans cet article, depuis l'extraction terminologique jusqu'à la formalisation en TEI-LMF.

1. Extraction terminologique augmentée
 - Constitution de deux corpus parallèles (français et espagnol) avec *SketchEngine* (Kilgariff *et al.*, 2004).
 - Identification des candidats termes via statistiques de fréquence / spécificité *keywords*, *term lists* et *word sketches*.
 - Filtrage complémentaire avec des modèles distributionnels (*fastText* ; Bojanowski *et al.*, 2017) entraînés sur nos corpus spécialisés.
 - Comparaison avec corpus généraux (*frWaC*, *esTenTen*) pour écarter les unités trop générales (ex. *système*, *projet*).
2. Alignement bilingue
 - Utilisation de modèles multilingues d'*embeddings* et de transfert non supervisé (*MUSE* ; Artetxe *et al.*, 2018).
 - Projection des vecteurs français / espagnols dans un espace commun, puis calcul de proximité cosinus.
 - Alignement automatique de paires comme par exemple *ossature bois* ↔ *estructura de madera* ou *béton précontraint* ↔ *hormigón pretensado*, ensuite validées manuellement.
3. Classification sémantique
 - Regroupement des unités en réseaux conceptuels avec des modèles contextuels de type *BERT* (Devlin *et al.*, 2019).
 - Modèles mobilisés : *CamemBERT* (Martin *et al.*, 2020) pour le français, *BETO* (Cañete *et al.*, 2020) pour l'espagnol.
 - Détection automatique de proximités comme *toiture végétalisée*, *toit vert* et *toiture extensive*, confirmées ensuite comme appartenant à une même hiérarchie conceptuelle.
4. Génération assistée de définitions et d'exemples
 - Extraction des contextes d'usage avec *SketchEngine*.
 - Condensation automatique par un modèle de langage (*GPT-4 – ChatGPT*) guidé par des *prompts* construits à partir des cooccurrences dominantes.
 - Exemple : pour *béton armé*, le modèle a mis en évidence les cooccurrences *résistance* et *armature métallique*, servant de base à une définition reformulée par le lexicographe.
5. Enrichissement sémantique contrôlé
 - Utilisation de modèles enrichis comme *SensEmBERT* (Scarlini *et al.*, 2020), combinant données distributionnelles et ressources lexicales

- préexistantes (*WordNet*, Miller, 1995 ; *BabelNet*, Navigli & Ponzetto, 2012).
- Suggestions semi-automatiques de relations : *toit vert* comme hyponyme de *toiture végétalisée*, ou rapprochement entre *ossature* et *estructura portante*.
 - Validation humaine systématique avant intégration.
6. Encodage TEI-LMF
- Normalisation des données lexicales selon TEI-LMF (ISO 24613).
 - Assistance de l'IA pour :
 - la pré-structuration des gabarits XML (<entry>, <form>, <sense>),
 - l'insertion d'attributs morphosyntaxiques (<gramGrp>, <pos>, <gen>),
 - l'intégration automatique des équivalents bilingues (<form type=«translation»>),
 - la détection de doublons et variantes via *embeddings* contextuels (*CamemBERT*, *BETO*) avec alerte si similarité cosinus > 0,85.
 - Génération semi-automatique de références croisées (<xr>) pointant vers les entrées concernées.
7. Supervision humaine
- Vérification et correction systématiques par les lexicographes.
 - Harmonisation conceptuelle, traitement des polysémies (ex. *module*), gestion des biais potentiels des modèles.

6. L'HUMAIN AU CŒUR DU PROCESSUS

L'analyse de notre démarche montre que l'intégration de l'IA dans le travail lexicographique exige en permanence une supervision humaine. Si l'automatisation a permis d'accélérer et de fiabiliser certaines étapes – extraction, alignement, classification, encodage –, la qualité finale de la ressource a reposé sur une validation experte menée de façon systématique par une personne.

Ce constat illustre un principe désormais reconnu dans les travaux récents sur l'e-lexicographie : l'IA agit comme un amplificateur de productivité, mais c'est l'humain qui en garantit la pertinence linguistique et conceptuelle (Lew, 2024 ; Li & Tarp, 2024).

Dans notre cas, plusieurs points ont illustré cette complémentarité :

- Valider les données générées : Les propositions générées par l'IA, qu'il s'agisse de définitions, d'exemples ou d'alignements, doivent être systématiquement vérifiés et validés par des experts du domaine. L'IA peut produire des erreurs, des ambiguïtés ou des informations non pertinentes qui nécessitent une correction humaine.
- Gérer les ambiguïtés et les polysémies : La langue, en particulier dans les domaines spécialisés, est souvent sujette à l'ambiguïté et à la polysémie.

L'IA peut avoir des difficultés à distinguer ces nuances sans une intervention humaine pour clarifier et spécifier les sens appropriés.

- Contrôler la cohérence et l'harmonisation : La cohérence terminologique est essentielle dans une ressource spécialisée. Les lexicographes assurent l'harmonisation des définitions, des relations sémantiques et du style, garantissant clarté et uniformité. L'IA peut signaler des incohérences, mais leur résolution dépend toujours de l'expertise humaine.
- Limiter les biais introduits : Les modèles d'IA sont entraînés sur des données existantes, ce qui peut introduire des biais présents dans ces données. La supervision humaine est nécessaire pour identifier et corriger ces biais, assurant ainsi que la ressource est équitable, inclusive et ne perpétue pas de stéréotypes ou de représentations erronées.
- Refléter l'évolution du domaine : L'architecture contemporaine est un domaine dynamique, avec l'émergence constante de nouveaux concepts, technologies et pratiques. Les lexicographes, en tant qu'experts du domaine, sont les mieux placés pour identifier ces évolutions et intégrer les nouveaux termes et concepts dans la ressource, assurant ainsi sa pertinence et son actualité.

7. CONCLUSION

Cette étude présente les principes de conception d'une ressource lexicale bilingue français-espagnol pour l'architecture contemporaine, un travail visant à combler une lacune significative dans les outils linguistiques spécialisés. En s'appuyant sur une modélisation en XML, conforme au standard LMF et intégrant une sérialisation TEI, nous avons posé les bases d'une ressource structurée, interopérable et pérenne.

Une dimension centrale de notre approche réside dans l'intégration stratégique de l'intelligence artificielle. En intégrant l'IA dans notre méthodologie, nous avons montré que la complémentarité entre automatisation et supervision humaine ouvre de nouvelles perspectives systématiques. Ainsi, l'IA n'est pas perçue comme un substitut à l'expertise humaine, mais comme un levier puissant pour automatiser des tâches lexicographiques répétitives telles que la génération de données, l'alignement terminologique et l'enrichissement contrôlé. Cette automatisation permet d'accroître l'efficacité du processus de création de la ressource, libérant ainsi les lexicographes pour des tâches à plus forte valeur ajoutée, nécessitant un jugement critique et une expertise conceptuelle.

Au-delà de ce projet, le workflow méthodologique que nous avons présenté peut également servir de feuille de route pour la conception d'autres ressources lexicales dans des domaines techniques variés, ou encore pour des langues sous-représentées, où les besoins en outils interopérables sont particulièrement importants.

BIBLIOGRAPHIE

- Artetxe, M., Labaka, G. & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In I. Gurevych & Y. Miyao (éds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (vol. 1 : Long Papers), 789-798. Melbourne : Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1073>
- Bartolomé-Díaz, Z. & Trujillo-González, V.C. (2023). Les termes architecturaux. Recommandations officielles et réalités des usages. *Cédille, revista de estudios franceses*, 24. <https://doi.org/10.25145/j.cedille.2023.24.12>
- Becchi, A. *et al.* (2008). Les livres d'architecture : Leurs éditions de la Renaissance à nos jours. *Perspective. Actualité en histoire de l'art*, 2. <https://doi.org/10.4000/perspective.3396>
- Bergholtz, H. & Tarp, S. (1995). *Manual of Specialised Lexicography*. Amsterdam (Pays-Bas) : John Benjamins Publishing Company. <https://benjamins.com/catalog/btl.12>
- Bojanowski, P. *et al.* (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146. https://doi.org/10.1162/tacl_a_00051
- Camacho-Collados, J. & Pilehvar, M.T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63 (1), 743-788. <https://doi.org/10.1613/jair.1.11259>
- Cañete, J. *et al.* (2020). Spanish Pre-trained BERT Model and Evaluation Data. Practical ML for Developing Countries Workshop : learning under limited/low resource scenarios. <https://doi.org/10.48550/arXiv.2308.02976>
- De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36. <https://doi.org/10.1093/ijl/ecad021>
- Devlin, J. *et al.* (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (vol. 1 : Long and Short Papers) pp. 4171-4186. Minneapolis, Minnesota (États-Unis) : Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Durand, B. (2023). *La fabrication d'une « architecture durable » en France (2000-2010)* [thèse de doctorat]. Université Paris-Est. <https://theses.hal.science/tel-04353097>
- Fung, J. *et al.* (2024). *Human-in-the-loop technical document annotation: Developing and validating a system to provide machine-assistance for domain-specific text analysis*. National Institute of Standards and Technology. <https://www.nist.gov/publications/human-loop-technical-document-annotation-developing-and-validating-system-provide>
- Grabowski, Ł. (2023). Statistician, programmer, data scientist? Who is, or should be, a corpus linguist in the 2020s? *Journal of Linguistics/Językovedný Casopis*, 74 (1), 52-59. <https://doi.org/10.2478/jazcas-2023-0023>
- Kilgarriff, A. *et al.* (2004). *The Sketch Engine*, 17.11.2016. 105-115. <https://euralex.org/publications/the-sketch-engine/>
- Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10 (1). <https://doi.org/10.1057/s41599-023-02119-6>
- Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11 (1), 1-8. <https://doi.org/10.1057/s41599-024-02889-7>
- L'Homme, M.-C. (2019). *Lexical semantics for terminology*. Amsterdam : John Benjamins Publishing Company.
- Li, Q. & Tarp, S. (2024). Using generative AI to provide high-quality lexicographic assistance to Chinese learners of English. *Lexikos*, 34, 397-418. <https://doi.org/10.5788/34-1-1944>

- Martin, L. *et al.* (2020). CamemBERT: A tasty French language model. In D. Jurafsky *et al.* (éds.), *En Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7203-7219). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.645>
- Mikolov, T. *et al.* (2013). Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations*. Scottsdale (AZ, États-Unis). <https://arxiv.org/abs/1301.3781>
- Miller, G.A. (1995). WordNet : A lexical database for English. In *Communications of the ACM*, vol. 38, (pp. 39-41). San Francisco (CA, États-Unis) : Morgan Kaufmann. <https://doi.org/10.1145/219717.219748>
- Navigli, R. & Ponzetto, S. (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>
- OpenAI *et al.* (2023). *GPT-4 Technical Report*. <https://cdn.openai.com/papers/gpt-4.pdf>
- Ortega-Martin, M. *et al.* (2023). *Spanish built factual freectionary (Spanish-BFF): The first AI-generated free dictionary*. <https://doi.org/10.48550/ARXIV.2302.12746>
- Ortega-Martin, M. *et al.* (2024). Building another Spanish dictionary, this time with GPT-4. <https://doi.org/10.48550/ARXIV.2406.11218>
- Periti, F., Alfter, D., & Tahmasebi, N. (2024). Automatically generated definitions and their utility for Modeling Word Meaning. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen (éds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 14008-14026). Miami, FL : Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.776>
- Romary, L. (2010). *Using the TEI framework as a possible serialization for LMF. Rendering endangered languages lexicons interoperable through standards harmonization*. RELISH Workshop Nijmegen (Pays-Bas). <https://hal.inria.fr/inria-00511769>
- Romary, L. (2013). TEI and LMF crosswalks. *Journal for Language Technology and Computational Linguistics* 30 (1), 47-70. <https://doi.org/10.21248/jlcl.30.2015.195>
- Romary, L. *et al.* (2019). LMF reloaded. In *Proceedings of the 13th International Conference of the Asian Association for Lexicography: Past, Present and Future* (pp. 533-539). Istanbul : Asos Publisher. <https://inria.hal.science/hal-02118319v1>
- Scarlini, B., Pasini, T. & Navigli, R. (2020). With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In B. Webber *et al.* (éds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 3528-3539). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.285>
- Toprak, A. & Turan, M. (2025). Automated thematic dictionary creation using the web based on WordNet, Spacy, and Simhash. *Data and Information Management*, 9 (3), 100088. <https://doi.org/10.1016/j.dim.2024.100088>
- Yu, W. *et al.* (2022). Dict-BERT : Enhancing language model pre-training with dictionary. Findings of the *Association for Computational Linguistics: ACL 2022*, 1907-1918. <https://doi.org/10.18653/v1/2022.findings-acl.150>