2026, vol. 6, n° Spécial INFORSID 2025, 98-111 pages, DOI : 10.21494/ISTE.OP.2026.1450 ISTE OpenScience

Caractérisation informatique et taxonomie des styles rédactionnels de brevets : vers une automatisation stylométrique de la rédaction technique

Computational Characterization and Taxonomy of Patent Writing Styles: Towards Stylometric Automation of Technical Drafting

Ghislain DEMONDA[1], Sébastien LABORIE[1], Christian SALLABERRY[1], Nathalie VALLES-PARLANGEAU[1]

[1] Université de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA, France

RÉSUMÉ. Dans le domaine de la propriété intellectuelle, les brevets sont des documents techniques et juridiques essentiels dont la rédaction requiert une expertise combinant des compétences techniques, juridiques et linguistiques. Les styles rédactionnels des brevets varient considérablement selon les domaines technologiques, juridictions et stratégies de protection. Cet article propose la conception de SCASB (Système de Caractérisation et d'Automatisation Stylométrique des Brevets), une approche unifiant pour la première fois les dimensions technique, juridique et stylistique dans un cadre informatique cohérent. Nous proposons une taxonomie bidimensionnelle des approches d'analyse de brevets (d'analyse automatique des documents × granularité d'analyse) et identifions les lacunes actuelles. Notre système se base sur l'évolution rapide des technologies de l'intelligence artificielle, particulièrement en traitement automatique du langage naturel (TALN). Ces travaux ouvrent la voie à une automatisation intelligente de la rédaction technique respectant les nuances stylistiques propres à chaque juridiction et stratégie de protection.

ABSTRACT. In the field of intellectual property, patents are essential technical and legal documents whose drafting requires expertise that combines technical, legal, and linguistic skills. Patent drafting styles vary considerably depending on technological domains, jurisdictions, and protection strategies. This article proposes the design of SCASB (System for Stylometric Characterization and Automation of Patents), an approach that, for the first time, unifies the technical, legal, and stylistic dimensions within a coherent computational framework. We propose a two-dimensional taxonomy of patent analysis approaches (automatic computational methods for document analysis × level of analytical granularity) and highlight current shortcomings. Our system builds upon the rapid advances in artificial intelligence technologies, particularly natural language processing (NLP). This work opens the path toward intelligent automation of technical drafting that accounts for the stylistic nuances specific to each jurisdiction and protection strategy.

MOTS-CLÉS. Analyse de brevets, Stylométrie, Traitement automatique du langage naturel, Apprentissage automatique, Rédaction technique.

KEYWORDS. Patent Analysis, Stylometry, Natural Language Processing, Machine Learning, Technical Writing.

1. Introduction

Les brevets constituent des documents techniques et juridiques fondamentaux pour la protection de l'innovation technologique. Leur rédaction nécessite une expertise multidisciplinaire rare, alliant compétences techniques, juridiques et linguistiques. Les styles rédactionnels varient considérablement selon les domaines technologiques, juridictions et stratégies de protection intellectuelle, posant des défis majeurs pour l'analyse, la classification et la génération automatique de ces documents.

Considérons le parcours typique de protection d'une innovation technologique : un inventeur décrit sa découverte à un ingénieur brevets lors d'un entretien technique. L'ingénieur brevets synthétise cette invention sous forme de revendications, c'est-à-dire des énoncés formels délimitant précisément la portée de la protection juridique accordée par le brevet. Les revendications constituent le cœur juridique et technique du document : c'est à partir d'elles que l'on détermine ce qui est protégé et ce qui ne l'est pas.

À partir de ces revendications, l'ingénieur brevets doit rédiger une description complète respectant les exigences formelles et stylistiques propres à la juridiction visée (Europe, États-Unis, Japon), au domaine technique (biotechnologie, électronique, mécanique) et à la stratégie de protection (défensive, offensive, de blocage). Les styles rédactionnels varient considérablement entre juridictions : les brevets américains privilégient les exemples concrets, les brevets européens favorisent l'abstraction, et les brevets japonais détaillent les variantes de l'invention revendiquée. La clarté et la fluidité rédactionnelle peuvent faciliter l'examen par les offices de brevets, bien que le contenu technique demeure identique.

Ce processus, actuellement largement manuel, pourrait bénéficier d'une assistance informatique. Imaginons un système capable d'analyser des milliers de brevets existants pour en extraire automatiquement les caractéristiques stylistiques, puis de générer ou d'adapter la rédaction d'un nouveau brevet en fonction du style approprié au contexte.

La stylométrie, discipline qui étudie le style d'écriture à l'aide de méthodes quantitatives, offre des outils prometteurs pour caractériser et classifier les styles rédactionnels des brevets. Cependant, son application au domaine spécifique des brevets reste largement inexplorée, en raison notamment de la complexité technique, des contraintes juridiques strictes et de la nature hautement normalisée de ces documents. La rédaction demeure une production humaine influencée par les préférences du rédacteur, permettant d'exprimer une même notion technique de multiples façons.

Notre recherche vise à répondre à la question fondamentale suivante : *Comment caractériser et classifier de manière informatique les styles rédactionnels des brevets pour permettre une automatisation stylométrique de la rédaction technique ?* À ce stade de nos travaux doctoraux (six mois), cet article pose les fondations conceptuelles du système SCASB. Nous présentons l'architecture théorique, les hypothèses de recherche et le protocole de validation envisagé. L'implémentation et l'évaluation empirique constitueront les phases ultérieures de notre recherche.

Cet article présente une approche pour l'automatisation stylométrique de la rédaction de brevets. La section 2 dresse un état de l'art structuré en deux dimensions : d'abord l'évolution des méthodes d'analyse automatique des documents (des règles manuelles à l'apprentissage profond), puis les applications spécifiques aux brevets (analyse stylométrique, classification et génération automatique). Cette analyse bidimensionnelle, synthétisée dans une taxonomie originale, permet d'identifier les lacunes actuelles et de positionner notre contribution. La section 3 présente notre système SCASB, détaillant son architecture technique, les hypothèses sous-jacentes et le protocole de validation. La section 4 discute les verrous scientifiques, techniques et épistémologiques à lever, propose une stratégie de résolution incrémentale et explore les extensions possibles à d'autres domaines documentaires au-delà des brevets.

2. État de l'art

L'analyse informatique des brevets s'articule autour de deux dimensions complémentaires : d'une part, les méthodes d'analyse automatique des documents employées (des règles manuelles à l'apprentissage profond), et d'autre part, les tâches visées (caractérisation stylistique, classification, génération).

Le Tableau 1 synthétise cette double perspective, par rapport à notre sujet de recherche, en croisant les niveaux d'automatisation (axe vertical) avec les niveaux d'analyse linguistique (axe horizontal : du lexique aux aspects pragmatiques et stratégiques).

La dimension pragmatique, telle que nous l'utilisons ici, désigne les aspects de la rédaction qui relèvent de l'intention communicative et du contexte d'usage du document, par opposition aux dimensions lexicale (choix des mots), syntaxique (structures grammaticales) et sémantique (sens des énoncés). Dans le cas des brevets, la dimension pragmatique recouvre notamment les stratégies de

protection (offensive, défensive, de blocage), les guides rédactionnels spécifiques à chaque juridiction et les conventions d'interaction avec les offices de brevets.

	Lexical	Syntaxique	Sémantique	Pragmatique
Règles manuelles	**Dictionnaires** *Termes techniques IPC*	**Grammaires** *Schémas de revendications*	**Ontologies** *Ontologie de brevets OMPI*	**Templates** *Guides rédactionnels*
Apprentissage automatique	**TF-IDF** *Classification CPC*	**CRF** *Extraction d'entités*	**LDA/LSA** *Clustering thématique*	**SVM** *Détection de stratégies*
Apprentissage profond	**Word2Vec** *Embeddings techniques*	**LSTM/GRU** *Génération de revendications*	**Modèles de langue** *PatentBERT, fine-tuning*	**GNN** *Analyse de citations*

Tableau 1. *Taxonomie bidimensionnelle - niveau d'automatisation (vertical) et granularité d'analyse (horizontal)*

Cette taxonomie présentée dans le Tableau 1 met en évidence une couverture inégale des dimensions d'analyse : si les niveaux lexical et sémantique bénéficient de travaux nombreux à tous les niveaux d'automatisation, la dimension stylistique, absente du tableau, n'a jamais fait l'objet d'une caractérisation systématique dans le contexte des brevets. C'est cette lacune que notre contribution vise à combler.

2.1. *Évolution des approches d'analyse automatique des documents*

2.1.1. *Approches basées sur des règles et l'expertise humaine*

Les premières tentatives d'analyse automatique des brevets s'appuient sur des approches symboliques exploitant l'expertise humaine formalisée sous forme de règles. Ces méthodes (cf. Masson [MASS 24]), bien que limitées dans leur capacité à traiter des cas non prévus par les règles, offrent l'avantage d'une transparence totale et d'une explicabilité directe.

Au niveau lexical, les approches à base de dictionnaires techniques constituent le fondement historique de l'analyse de brevets. Les systèmes de classification IPC (International Patent Classification) et CPC (Cooperative Patent Classification) reposent sur des taxonomies hiérarchiques de termes techniques soigneusement élaborées par des experts. Ces classifications opèrent sur le contenu technique des brevets et ne portent pas sur leur style rédactionnel (Fall et coll. [FAL 03] ; [WIPO 22]). Ces dictionnaires permettent une catégorisation précise, mais rigide, nécessitant des mises à jour manuelles régulières pour suivre l'évolution technologique.

Sur le plan syntaxique, des grammaires formelles ont été proposées pour capturer les structures récurrentes des revendications de brevets (voir [KRE 21] pour un panorama de l'évolution des méthodes d'analyse de brevets). Ces schémas linguistiques, tels que « comprenant » (« comprising » ou « wherein » en anglais), « caractérisé en ce que » (« characterized by » en anglais), forment un langage quasi-formel que les systèmes à base de règles peuvent exploiter pour l'extraction d'informations structurées.

Au niveau sémantique, des ontologies de brevets ont été développées pour capturer les relations conceptuelles entre inventions, domaines techniques et concepts juridiques [WIPO 22]. Ces

représentations formelles de la connaissance permettent des inférences logiques, mais souffrent de la difficulté à constituer et maintenir ces ontologies.

Au niveau pragmatique, des guides rédactionnels institutionnels, tels que le WIPO Patent Drafting Manual [WIPO 23], formalisent les conventions d'écriture propres à chaque juridiction et fournissent des templates structurels pour les différentes sections d'un brevet. Ces ressources constituent l'état actuel de la formalisation du savoir-faire rédactionnel, mais elles restent descriptives et ne permettent pas une exploitation computationnelle directe.

2.1.2. *Méthodes d'apprentissage automatique classique*

L'introduction des méthodes statistiques d'apprentissage automatique a marqué un tournant dans l'analyse automatique des brevets, permettant de dépasser les limitations des approches purement symboliques. Ces méthodes se déclinent en approches supervisées et non supervisées.

Les techniques de vectorisation comme TF-IDF (Term Frequency-Inverse Document Frequency) ont permis les premières classifications thématiques automatiques efficaces de brevets selon les codes IPC et CPC. Ces représentations vectorielles, bien que perdant la structure syntaxique, capturent efficacement la distribution thématique des documents. Les travaux pionniers de Fall et coll. [FAL 03], utilisant quatre algorithmes de classification (Naïve Bayes, k-NN, SVM) sur des représentations vectorielles des représentations bag-of-words des documents, rapportent, sur les 300 premiers mots des documents, des précisions de 51 à 79% au niveau des classes IPC (114 classes) selon l'algorithme et la mesure employée, les meilleurs résultats étant atteints avec la mesure « three-guesses » (73 à 79%). Au niveau des sous-classes (451 sous-classes), les performances diminuent significativement, avec un top-1 de 33 à 41% et un three-guesses de 53 à 62% dans la configuration standard, pouvant atteindre 65% avec un entraînement sur des documents à sous-classe unique.

Les modèles de séquences comme les CRF (Conditional Random Fields, Lafferty et coll. [LAF 01]) ont apporté une solution au problème de l'extraction d'entités nommées dans les brevets, notamment pour l'identification d'entités chimiques dans les documents de propriété industrielle (Grego et coll. [GRE 09]). Leur capacité à modéliser les dépendances contextuelles les rend particulièrement adaptés à la nature hautement structurée du langage des brevets.

Les machines à vecteurs de support (SVM, Cortes et Vapnik [COR 95]) ont démontré leur efficacité pour des tâches de classification binaire ou multi-classe complexes dans le domaine des brevets, notamment la prédiction de la valeur économique des brevets (Ercan et Kayakutlu [ERC 14]). Leur capacité à gérer des espaces de caractéristiques de haute dimension les rend particulièrement adaptés à l'analyse textuelle des brevets.

Parallèlement à ces approches supervisées, des méthodes non supervisées ont ouvert de nouvelles perspectives pour l'analyse de corpus de brevets.

Les approches de modélisation thématique non supervisées, notamment LSA (Latent Semantic Analysis, Deerwester et coll. [DEE 90]) et LDA (Latent Dirichlet Allocation, Blei et coll. [BLE 03]) ont ouvert la voie à l'analyse sémantique latente des corpus de brevets. Ces techniques révèlent des structures thématiques cachées, permettant le clustering automatique de brevets similaires sans supervision humaine. LDA en particulier s'est révélé efficace pour identifier les tendances technologiques émergentes dans de larges corpus de brevets (Kim et Lee [KIM 15]).

2.1.3. *Architectures d'apprentissage profond*

L'analyse informatique des brevets bénéficie aujourd'hui des avancées en apprentissage profond, particulièrement dans le domaine du traitement automatique du langage naturel. Les principales architectures explorées sont les transformeurs et les réseaux de neurones graphiques (GNN).

Les modèles d'embeddings comme Word2Vec (Mikolov et coll. [MIK 13]) et FastText (Bojanowski et coll. [BOJ 17]) ont révolutionné la représentation vectorielle du vocabulaire technique des brevets. Ces représentations denses capturent les similarités sémantiques entre termes techniques, permettant par exemple de reconnaître que « transistor » et « semi-conducteur » appartiennent au même champ sémantique. Ces embeddings servent de fondation pour de nombreuses tâches en aval.

Les architectures récurrentes LSTM (Long Short-Term Memory, Hochreiter et Schmidhuber [HOC 97]) et GRU (Gated Recurrent Unit) ont été explorées pour diverses tâches d'analyse et de génération de textes de brevets (cf. Krestel et coll. [KRE 21]). Leur capacité à modéliser les dépendances à long terme est particulièrement adaptée aux phrases complexes et enchâssées typiques du langage des brevets.

Krestel et coll. [KRE 21] proposent une revue des applications de l'apprentissage profond pour l'analyse des brevets, soulignant le potentiel des modèles transformeurs pour améliorer la classification et l'extraction d'informations. Lee et coll. [LEE 19] présentent PatentBERT, un modèle adapté aux brevets qui améliore les performances de 15% par rapport aux approches traditionnelles. Ces modèles exploitent le mécanisme d'attention pour capturer les relations complexes entre concepts techniques distants dans le texte.

Ding et coll. [DIN 24] utilisent des réseaux de neurones graphiques (GNN, Graph Neural Networks) pour modéliser la dynamique des citations de brevets. Cette approche traite le réseau de citations comme un graphe dynamique, permettant de prédire les trajectoires futures de citation et d'identifier les brevets potentiellement influents.

2.2. *Travaux intégrant la dimension stylistique*

Les approches présentées en section 2.1 se concentrent principalement sur la caractérisation thématique et technique des brevets (classification par domaine, extraction d'entités, modélisation de sujets). La dimension stylistique de ces documents, c'est-à-dire la manière dont un même contenu technique est rédigé selon les contextes, reste en revanche largement sous-explorée. Quelques travaux abordent néanmoins cet aspect, que ce soit dans le contexte des brevets ou dans des contextes proches. Nous les examinons ci-dessous), avant d'identifier les lacunes actuelles et de positionner notre contribution (section 2.3).

Si les méthodes d'analyse stylistique sont bien établies en littérature (cf. Stamatatos [STA 09]), leur application aux brevets reste limitée en raison d'obstacles spécifiques : inadaptation des métriques stylistiques traditionnelles aux structures hautement normalisées, complexité d'extraction des caractéristiques sur des sections hétérogènes (description technique, revendications juridiques, dessins) et multilingues, et cloisonnement entre expertises techniques, juridiques et linguistiques.

Parmi les travaux qui abordent néanmoins la dimension stylistique, Gehrmann et coll. [GEH 19] ont développé GLTR (Giant Language model Test Room), un outil statistique pour l'analyse et la détection de texte généré automatiquement. Bien que non spécifiquement conçu pour les brevets, GLTR offre des éclairages sur les schémas stylistiques distinguant rédaction humaine et génération automatique, une distinction pertinente dans le contexte juridique des brevets où l'authenticité est primordiale.

Dans le contexte spécifique des brevets, Chen et Zhu [CHE 20] proposent une approche de transfert de style à faibles données (few-shot learning) pour les documents juridiques. Leur méthode, fondée sur une architecture de méta-apprentissage, démontre qu'il est possible d'adapter le style rédactionnel avec seulement quelques exemples d'entraînement, ce qui représente un avantage considérable dans un domaine où les corpus annotés sont rares et coûteux.

Du côté de la génération, Casola et coll. [CAS 22] explorent l'application de l'ingénierie d'instructions (« prompt engineering ») pour la génération de différentes sections de brevets. Leur approche démontre que des instructions soigneusement conçues peuvent guider les modèles de langage vers la production de textes respectant les conventions stylistiques spécifiques à chaque section. Cependant, ces modèles

restent limités dans leur capacité à contrôler finement la variation stylistique selon les paramètres contextuels (juridiction, domaine, stratégie).

2.3. *Lacunes identifiées et positionnement*

L'analyse approfondie de l'état de l'art révèle plusieurs verrous scientifiques qui justifient notre recherche et définissent notre positionnement.

Premièrement, malgré l'existence de cadres de classification technique bien établis comme l'IPC ou la CPC, la dimension stylistique de la rédaction des brevets n'a jamais fait l'objet d'une caractérisation systématique. Les guides pratiques disponibles, notamment le WIPO Patent Drafting Manual [WIPO 23] et le WIPO Patent Analytics Handbook [WIPO 22], fournissent certes des recommandations utiles et des cartographies de codes techniques, mais ils ne proposent pas d'analyse formelle des variations stylistiques aux niveaux micro-textuel (choix lexicaux, structures syntaxiques, marqueurs discursifs au sein des phrases et paragraphes) et macro-textuel (organisation globale du document, séquençage des sections, stratégies argumentatives d'ensemble). Cette absence de caractérisation stylistique constitue un frein majeur à l'automatisation de la rédaction de brevets.

Deuxièmement, aucun modèle existant n'intègre simultanément les dimensions technique, juridique et stylistique des brevets dans un cadre informatique unifié. Les approches actuelles traitent ces dimensions de manière isolée : les modèles techniques ignorent les contraintes juridiques, les systèmes juridiques négligent les nuances techniques, et les approches stylistiques ne considèrent pas les spécificités du domaine des brevets.

Troisièmement, l'absence de taxonomies adaptées aux spécificités rédactionnelles transjuridictionnelles empêche toute comparaison systématique et toute génération contrôlée. Les différences stylistiques entre un brevet américain (privilégiant les exemples concrets), européen (favorisant l'abstraction) et japonais (détaillant les variantes) ne sont pas formellement modélisées.

Quatrièmement, les approches actuelles de génération automatique ne permettent pas de moduler le style selon les contraintes contextuelles. Un même contenu technique devrait pouvoir être exprimé différemment selon qu'il s'agit d'une stratégie défensive (large couverture) ou offensive (revendications précises), mais les systèmes actuels ne capturent pas ces nuances stratégiques.

Notre approche SCASB vise précisément à combler ces lacunes en proposant un système intégré de caractérisation et de génération stylométrique qui unifie les dimensions technique, juridique et stylistique dans un cadre informatique cohérent.

3. Architecture proposée : le système SCASB

3.1. *Vue d'ensemble du système*

Le système SCASB que nous proposons se distinguerait des approches existantes par sa capacité à :

- extraire automatiquement des profils stylistiques multidimensionnels à partir de corpus de brevets existants ;
- classifier ces profils selon trois axes : juridictionnel, technique et stratégique ;
- générer ou adapter des textes de brevets en respectant le style approprié au contexte cible (fonctionnalité prospective).

Notre recherche vise à développer des méthodes capables de capturer les subtilités linguistiques, techniques et juridiques qui définissent le style rédactionnel d'un brevet, tout en restant adaptables aux différents contextes. Nous proposons l'élaboration d'une taxonomie des styles rédactionnels de brevets

intégrant les variations liées aux domaines technologiques, juridictions et stratégies de protection intellectuelle.

La Figure 1 présente l'architecture globale du système SCASB. Le système s'articule en quatre modules fonctionnels orchestrés selon trois étapes de traitement séquentielles (détaillées en section 3.2).

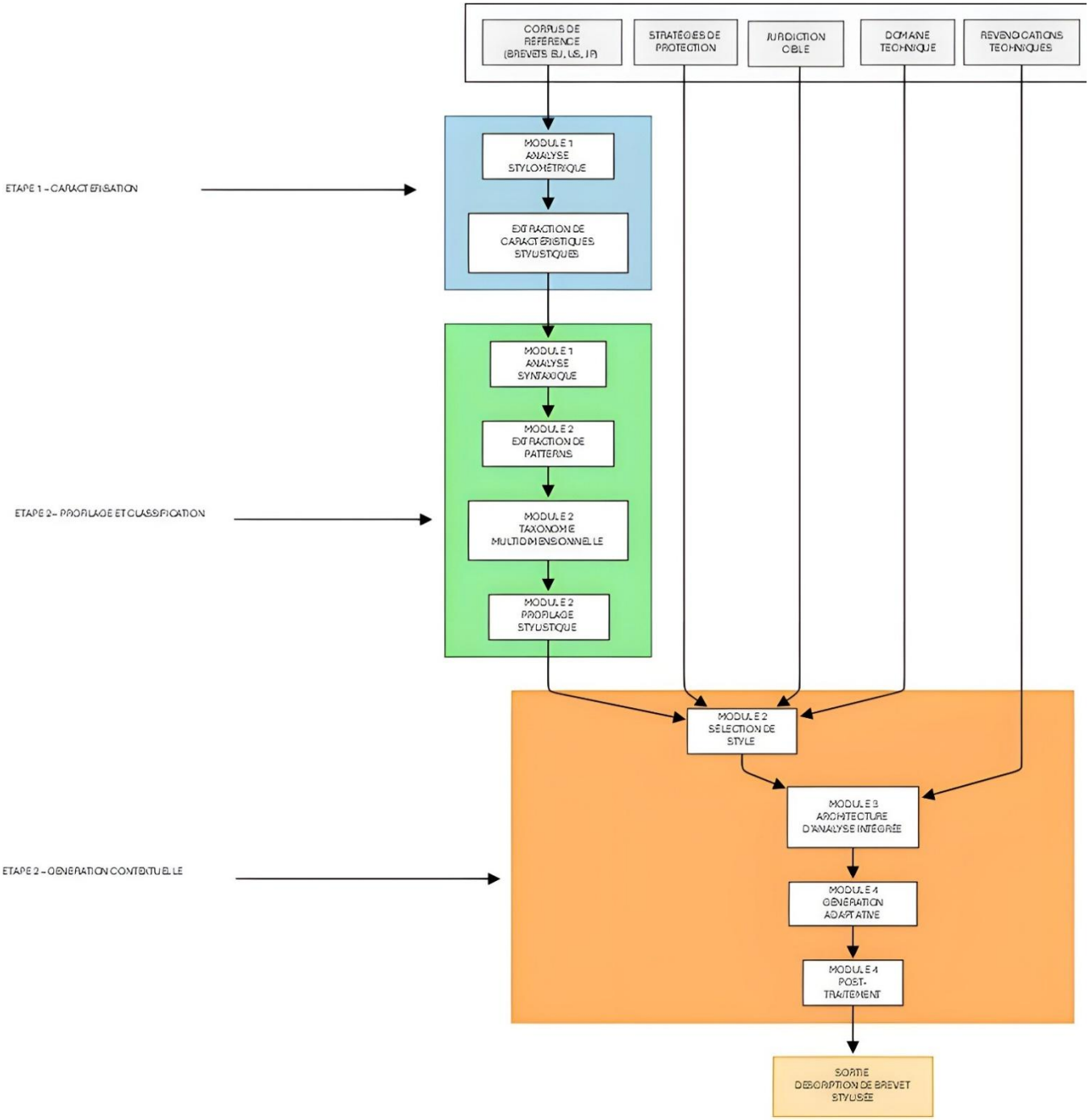

Figure 1. *Architecture du système SCASB. En entrée : revendications techniques et paramètres contextuels (juridiction, domaine, stratégie). En sortie : description de brevet stylisée selon les paramètres.*

Le système s'articule en quatre modules fonctionnels (Modules 1-4) orchestrés selon trois phases de traitement séquentielles (indiquées par les zones colorées dans le diagramme).

Le Module 1 extrait les caractéristiques stylométriques du corpus de référence. Le Module 2 organise ces caractéristiques dans une taxonomie multidimensionnelle. Les Modules 3 et 4 génèrent la description du brevet en appliquant le profil stylistique approprié aux revendications techniques fournies en entrée.

Cette architecture intègre les trois dimensions identifiées comme essentielles : technique (Module 1), juridique (Module 2), et stylistique (Modules 3 et 4).

3.2. *Architecture technique, hypothèses et validation*

3.2.1. *Hypothèses de travail*

Les méthodes traditionnelles de stylométrie n'étant pas directement applicables aux brevets en raison de leur nature hautement structurée et normalisée, notre recherche se concentre sur la conception de nouveaux algorithmes capables de traiter efficacement les particularités de ces documents. Nous formulons trois hypothèses principales qui guident notre approche :

H1 - Caractérisation multidimensionnelle : les styles rédactionnels de brevets peuvent être caractérisés par des vecteurs de caractéristiques combinant des métriques lexicales (terminologie technique spécifique au domaine, expressions juridiques standardisées), syntaxiques (longueur et complexité des phrases, structures enchâssées typiques des revendications) et structurelles (organisation hiérarchique des sections, schémas de référencement interne et externe).

H2 - Supériorité des approches neuronales : l'utilisation de modèles de langage préentraînés (architectures de modèles de langue) avec fine-tuning sur des corpus spécialisés de brevets permet de capturer les nuances stylistiques subtiles mieux que les approches purement statistiques, grâce à leur capacité à modéliser les dépendances contextuelles longues et les relations sémantiques implicites.

H3 - Viabilité économique par apprentissage few-shot : un apprentissage few-shot permet de transférer efficacement le style entre juridictions avec seulement quelques exemples (5-10 brevets), rendant le système économiquement viable pour des domaines techniques émergents ou des juridictions moins documentées.

3.2.2. *Workflow de traitement : trois étapes séquentielles*

Les quatre modules fonctionnels (présentés en Figure 1 et détaillés en section 3.2.3) s'orchestreraient selon un workflow séquentiel en trois phases lors du traitement d'une demande de génération de brevet :

Étape 1 - Caractérisation (Module 1) : Les brevets du corpus de référence (de nombreux documents brevets issus des bases USPTO, EPO et JPO) sont analysés par le module de caractérisation stylométrique. Cette analyse produit une représentation vectorielle multidimensionnelle de chaque brevet, capturant ses caractéristiques stylistiques à différents niveaux de granularité.

Étape 2 - Profilage et classification (Modules 1 et 2) : Les représentations vectorielles sont ensuite organisées dans l'espace stylistique tridimensionnel (juridiction × domaine × stratégie) par des techniques de clustering non supervisé et de classification supervisée. Cette étape génère une bibliothèque de profils stylistiques réutilisables.

Étape 3 - Génération contextuelle (Modules 2, 3 et 4) : Pour générer un nouveau brevet, le système reçoit en entrée les revendications techniques et les paramètres contextuels (juridiction cible, domaine technique, stratégie de protection). Le module de génération sélectionne le profil stylistique approprié dans la bibliothèque et l'applique pour produire une description de brevet conforme aux conventions identifiées.

3.2.3. *Architecture modulaire et verrous scientifiques*

Le système SCASB s'articulerait autour de quatre composants principaux interconnectés, chacun présentant des défis scientifiques et techniques spécifiques que notre approche vise à surmonter à travers les trois étapes du workflow (section 3.2.2).

Module 1 - Analyse stylométrique de traitement automatique des brevets : ce module extrait et quantifie les caractéristiques stylistiques à plusieurs niveaux de granularité. Au niveau lexical, il calcule

la distribution des termes techniques, la densité de jargon juridique et les schémas de néologismes. Au niveau syntaxique, il analyse la complexité des structures grammaticales, les schémas de subordination et les marqueurs discursifs spécifiques aux brevets.

Verrous scientifiques du Module 1 :

- Hétérogénéité des corpus : les brevets varient de 10 à 500 pages selon les domaines techniques, avec des structures et des niveaux de complexité très disparates. Comment normaliser l'extraction de caractéristiques sur des documents aussi hétérogènes ?

- Multilinguisme juridique : la traduction des concepts entre systèmes juridiques différents (common law vs droit civil) dépasse le cadre de la simple traduction linguistique et nécessite une compréhension profonde des implications légales dans chaque juridiction.

- Coût d'annotation : la constitution de jeux de données annotés se heurte à des contraintes économiques, l'expertise requise pour annoter stylométriquement des brevets rendant ce processus particulièrement coûteux.

Module 2 - Taxonomie multidimensionnelle : ce composant organise les styles identifiés selon trois dimensions orthogonales. La dimension juridictionnelle capture les différences entre les conventions rédactionnelles américaines (orientées exemples), européennes (privilégiant l'abstraction) et asiatiques (détaillant les variantes). La dimension technique distingue les styles propres aux domaines (biotechnologie vs électronique vs mécanique). La dimension stratégique différencie les approches défensives (couverture large) des approches offensives (revendications précises).

Verrous scientifiques du Module 2 :

- Absence de taxonomies transjuridictionnelles : les différences stylistiques entre brevets américains, européens et japonais ne sont pas formellement modélisées, empêchant toute comparaison systématique et toute génération contrôlée.

- Définition du style en contexte juridico-technique : où tracer la frontière entre les contraintes formelles imposées par les offices de brevets et les véritables choix stylistiques du rédacteur ? La mesure de la qualité rédactionnelle soulève des interrogations similaires, les critères variant selon les acteurs (inventeurs cherchant la clarté, examinateurs privilégiant la précision, juges valorisant l'exhaustivité).

Module 3 - Architecture d'analyse intégrée : l'architecture technique combine des transformeurs pour l'analyse sémantique profonde avec des réseaux de neurones graphiques (GNN) pour modéliser les relations structurelles entre sections du brevet. Cette approche hybride permet de capturer à la fois le contenu sémantique et la structure documentaire, deux aspects essentiels des brevets.

Verrous scientifiques du Module 3 :

- Évolutivité : le traitement de millions de brevets nécessite le développement d'architectures distribuées et l'implémentation d'optimisations algorithmiques sophistiquées.

- Interprétabilité : les praticiens du droit exigent légitimement de comprendre les mécanismes sous-jacents aux recommandations stylistiques proposées, posant le défi de l'explicabilité des décisions du système.

Module 4 - Système de génération adaptative : le module de génération utilise une architecture encoder-decoder augmentée d'un mécanisme de contrôle stylistique. L'encodeur traite les revendications techniques d'entrée, tandis que le décodeur, conditionné par les paramètres stylistiques cibles, génère la description adaptée. Un mécanisme d'attention croisée assure la cohérence entre le contenu technique et le style appliqué.

Verrous scientifiques du Module 4 :

- Confidentialité : le traitement de brevets en cours de dépôt impose des protocoles de sécurité renforcés pour protéger la propriété intellectuelle des inventeurs.

- Standardisation vs diversité : l'éthique de l'automatisation soulève des questions sur l'impact sur la profession d'ingénieur brevets et les risques d'une standardisation excessive qui pourrait nuire à la richesse des stratégies de protection intellectuelle.

Ces verrous seront levés progressivement à travers les trois étapes du workflow décrites en section 3.2.2 : l'Étape 1 (caractérisation) adresse les verrous du Module 1, l'Étape 2 (profilage) traite les verrous du Module 2, et l'Étape 3 (génération) résout les verrous des Modules 3 et 4.

3.2.4. *Métriques d'évaluation et protocole de validation*

Pour valider rigoureusement notre approche, nous proposons un protocole d'évaluation à trois niveaux complémentaires :

1. **Évaluation automatique quantitative** :
    - Mesures de similarité stylistique utilisant la distance de Jensen-Shannon entre distributions lexicales pour quantifier la fidélité au style cible
    - Scores BLEU et ROUGE pour évaluer la qualité de génération par rapport à des brevets de référence
    - Métriques de cohérence interne (perplexité, cohérence thématique) pour assurer la qualité textuelle
    - Analyse de la couverture conceptuelle pour vérifier que tous les aspects techniques des revendications sont traités
2. **Évaluation par experts du domaine** :
    - Validation en double aveugle par plusieurs ingénieurs brevets qualifiés de la conformité juridique et de l'adéquation stylistique
    - Étude comparative sur plusieurs brevets (certains seront générés vs certains seront rédigés manuellement) pour évaluer l'indistinguabilité
    - Questionnaire structuré évaluant : clarté technique (échelle 1-5), conformité juridique (binaire), adéquation stylistique (échelle 1-5)
    - Analyse qualitative des commentaires d'experts pour identifier les points d'amélioration
3. **Évaluation applicative :**
    - Mesure du temps de rédaction : réduction attendue avec assistance SCASB versus rédaction entièrement manuelle
    - Taux d'acceptation par les offices de brevets : comparaison du taux de rejets formels entre brevets assistés et manuels
    - Satisfaction utilisateur : enquête auprès des utilisateurs professionnels sur l'utilisabilité et la valeur ajoutée perçue

- ROI économique : évaluation du rapport coût/bénéfice incluant les coûts de formation et d'intégration

Ce protocole d'évaluation multi-facettes, qui n'existe pas actuellement dans la littérature pour l'évaluation de systèmes de génération de brevets, permettra de valider non seulement la performance technique du système, mais aussi son utilité pratique et sa viabilité économique. Nous proposons ce cadre méthodologique comme contribution pour standardiser les évaluations comparatives entre différentes approches et systèmes dans le domaine.

4. Discussion et perspectives

4.1. *Stratégies de résolution envisagées*

Pour lever les verrous identifiés en section 3.2.3, le système SCASB s'appuiera sur les trois étapes du workflow présentées en section 3.2.2. Nous proposons une approche incrémentale structurée en trois phases progressives correspondant chacune au développement et à la validation d'une étape du workflow (section 3.2.2).

La première phase visera à établir une preuve de concept solide en constituant un corpus pilote de plusieurs brevets sélectionnés dans trois domaines techniques représentatifs. Ce volet permettra de développer un prototype initial de l'étape 1 (caractérisation stylométrique) et de le valider sur un cas d'usage restreint, mais significatif, à savoir l'adaptation de brevets européens aux exigences stylistiques américaines.

La deuxième pahse permettra d'élargir considérablement le périmètre en intégrant dix domaines techniques et cinq juridictions majeures, développant ainsi l'étape 2 (profilage et classification). L'intégration de techniques d'apprentissage few-shot constituera un axe de développement prioritaire pour réduire les besoins en données annotées, répondant ainsi aux contraintes économiques identifiées. Des études utilisateurs seront menées avec des ingénieurs brevets partenaires pour valider l'utilisabilité et la pertinence des outils développés.

La troisième phase se concentrera sur le déploiement de l'étape 3 (génération contextuelle) en conditions réelles et l'évaluation d'impact. Des tests seront conduits dans des cabinets de propriété intellectuelle partenaires pour mesurer concrètement les gains de productivité et l'amélioration de la qualité rédactionnelle. Une analyse des implications socio-économiques de l'automatisation sera également menée pour anticiper et accompagner les transformations de la profession.

4.2. *Applications et extensions potentielles*

Cette recherche adresse la caractérisation informatique des styles rédactionnels de brevets en s'articulant autour de quatre axes complémentaires. Les contributions scientifiques attendues comprennent une méthodologie robuste d'analyse des styles, une taxonomie validée empiriquement, une architecture d'analyse stylométrique spécialisée et un système d'aide à la rédaction.

Sur le plan applicatif, nos travaux permettront d'améliorer la qualité rédactionnelle des brevets, de réduire les coûts de rédaction, d'assurer une meilleure conformité aux exigences juridiques et de fournir des outils d'aide à la décision.

Les perspectives d'extension de l'approche SCASB dépassent largement le cadre des brevets. Le système pourrait naturellement s'adapter à d'autres types de documents juridiques tels que les contrats internationaux, les décisions de justice ou les textes réglementaires, chacun présentant ses propres défis stylistiques et contraintes formelles.

Dans le domaine de la documentation technique, l'approche pourrait s'appliquer aux manuels d'utilisation, aux spécifications techniques et aux normes industrielles, facilitant leur adaptation à différents publics et contextes d'usage.

La rédaction scientifique constitue également un champ d'application prometteur, notamment pour l'adaptation stylistique d'articles selon les exigences spécifiques des revues cibles.

L'intégration future de modalités complémentaires, incluant les schémas techniques, les formules chimiques et les équations mathématiques, enrichirait considérablement les capacités du système. Parallèlement, l'application de techniques d'explicabilité avancées, telles que les cartes d'attention ou les valeurs SHAP, améliorerait la transparence et favoriserait l'acceptabilité du système auprès des professionnels du domaine.

5. Conclusion

Les brevets constituent des documents techniques et juridiques dont la complexité stylistique reste largement inexplorée par les approches informatiques existantes. Leur rédaction nécessite une expertise rare combinant compétences techniques, juridiques et linguistiques, et les styles varient considérablement selon les juridictions, domaines technologiques et stratégies de protection. Cet article a proposé le cadre conceptuel du système SCASB, qui unifie pour la première fois les dimensions technique, juridique et stylistique dans une architecture informatique cohérente.

Nos contributions se situent à trois niveaux. Premièrement, nous avons élaboré une taxonomie bidimensionnelle des approches d'analyse de brevets croisant les méthodes d'analyse automatique (règles manuelles, apprentissage supervisé/non supervisé, apprentissage profond) avec les granularités d'analyse (lexicale, syntaxique, sémantique, pragmatique). Cette taxonomie révèle que les dimensions stylistique et pragmatique demeurent sous-explorées, particulièrement aux niveaux d'automatisation élevés. Deuxièmement, nous avons conçu une architecture modulaire intégrant quatre composants fonctionnels : analyse stylométrique, taxonomie multidimensionnelle, architecture d'analyse intégrée et système de génération adaptative. Ces modules s'orchestrent selon un workflow en trois étapes permettant la caractérisation, le profilage et la génération de brevets. Troisièmement, nous avons défini un protocole de validation multi-niveaux combinant évaluation automatique quantitative, évaluation par experts du domaine et évaluation applicative sur des métriques de productivité et de qualité rédactionnelle.

À ce stade de notre recherche doctorale, ces travaux conceptuels établissent les fondations nécessaires aux développements futurs. Les phases suivantes comprendront la constitution d'un corpus annoté multilingue couvrant plusieurs juridictions et domaines techniques, l'implémentation des modules de caractérisation et de génération basés sur des architectures de transformeurs spécialisées, et la validation empirique avec des ingénieurs brevets dans des conditions réelles d'utilisation. Les trois hypothèses formulées (caractérisation multidimensionnelle, supériorité des approches neuronales, viabilité de l'apprentissage few-shot) devront être testées empiriquement pour évaluer la pertinence de notre approche.

Au-delà du domaine des brevets, cette recherche explore une question scientifique plus large : comment caractériser et automatiser le style dans des genres documentaires hautement normalisés où contraintes formelles imposées par les institutions et choix rédactionnels individuels s'entremêlent ? Les méthodologies développées pourraient s'étendre à d'autres types de documents juridiques, à la documentation technique normalisée, ou à la rédaction scientifique soumise à des conventions éditoriales strictes. L'intégration future de modalités complémentaires (schémas techniques, formules chimiques, équations mathématiques) et l'application de techniques d'explicabilité avancées constitueront des axes de développement prometteurs pour renforcer l'acceptabilité et la transparence du système auprès des professionnels du domaine.

Remerciements

Nous remercions le Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour (LIUPPA) pour son soutien.

Bibliographie

[BERR 23] BERRIOS TORRES A., Language models for PATENTS: exploring prompt engineering for the patent domain, Mémoire de master, Politecnico di Milano, 2023.

[BLE 03] BLEI D. M., NG A. Y., JORDAN M. I., « Latent Dirichlet allocation », Journal of Machine Learning Research, vol. 3, p. 993-1022, 2003.

[BOJ 17] BOJANOWSKI P., GRAVE E., JOULIN A., MIKOLOV T., « Enriching word vectors with subword information », Transactions of the Association for Computational Linguistics, vol. 5, p. 135-146, 2017.

[CAS 22] CASOLA S., LAVELLI A., « Summarization, simplification, and generation: the case of patents », Expert Systems with Applications, vol. 205, 117627, 2022.

[CHE 20] CHEN X., ZHU K. Q., « Small-data text style transfer via multi-task meta-learning », https://arxiv.org/abs/2004.11742, 2020.

[COR 95] CORTES C., VAPNIK V., « Support-vector networks », Machine Learning, vol. 20, n° 3, p. 273-297, 1995.

[DEE 90] DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K., HARSHMAN R., « Indexing by latent semantic analysis », Journal of the American Society for Information Science, vol. 41, n° 6, p. 391-407, 1990.

[DIN 24] DING M., YU W., ZENG T., WANG S., « PTNS: patent citation trajectory prediction based on temporal network snapshots », Scientific Reports, vol. 14, art. 24034, 2024.

[ERC 14] ERCAN S., KAYAKUTLU G., « Patent value analysis using support vector machines », Soft Computing, vol. 18, n° 2, p. 313-328, 2014.

[FAL 03] FALL C. J., TÖRCSVÁRI A., BENZINEB K., KARETKA G., « Automated categorization in the international patent classification », ACM SIGIR Forum, vol. 37, n° 1, p. 10-25, 2003.

[GEH 19] GEHRMANN S., STROBELT H., RUSH A. M., « GLTR: statistical detection and visualization of generated text », in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, p. 111-116, 2019.

[GRE 09] GREGO T., PĘZIK P., COUTO F. M., REBHOLZ-SCHUHMANN D., « Identification of chemical entities in patent documents », in: IWANN 2009, Part II, LNCS, vol. 5518, p. 942-949, Springer, 2009.

[HOC 97] HOCHREITER S., SCHMIDHUBER J., « Long short-term memory », Neural Computation, vol. 9, n° 8, p. 1735-1780, 1997.

[KIM 15] KIM J., LEE J.-H., « Technology analysis from patent data using latent Dirichlet allocation », in : Proceedings of the Portland International Conference on Management of Engineering and Technology (PICMET), 2015.

[KRE 21] KRESTEL R., CHIKKAMATH R., HEWEL C., RISCH J., « A survey on deep learning for patent analysis », World Patent Information, vol. 65, 102035, 2021. DOI : 10.1016/j.wpi.2021.102035.

[LAF 01] LAFFERTY J., MCCALLUM A., PEREIRA F., « Conditional random fields: probabilistic models for segmenting and labeling sequence data », in: Proceedings of the 18th International Conference on Machine Learning (ICML), p. 282-289, 2001, https://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf.

[LEE 19] LEE J., et al., « PatentBERT : patent classification with fine-tuning a pre-trained BERT model », https://arxiv.org/abs/1906.02124, 2019.

[MASS 24] MASSON M., « Generic framework for the multidimensional processing and analysis of social media content, a proxemic approach », Thèse de doctorat, Université de Pau et des Pays de l'Adour, 2024.

[MIK 13] MIKOLOV T., CHEN K., CORRADO G., DEAN J., « Efficient estimation of word representations in vector space », in : Proceedings of the 1st International Conference on Learning Representations (ICLR 2013), Workshop Track, arXiv:1301.3781, 2013.

[STA 09] STAMATATOS E., « A survey of modern authorship attribution methods», Journal of the American Society for Information Science and Technology, vol. 60, n° 3, p. 538-556, 2009.

[WIPO 22] WORLD INTELLECTUAL PROPERTY ORGANIZATION (WIPO), WIPO Patent Analytics Handbook, Geneva, WIPO, 2022, https://wipo-analytics.github.io/handbook/.

[WIPO 23] WORLD INTELLECTUAL PROPERTY ORGANIZATION (WIPO), WIPO Patent Drafting Manual, 2nd edition, Geneva, WIPO, 2023, https://www.wipo.int/publications/en/details.jsp?id=4706.