Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Mise en correspondance de nomenclatures métiers à partir d'une extraction et d'un alignement des tâches décrites

Mathilde Geley, Estelle Laurence, Gatien Leclercq, Louve Maestre, Anaïs Schlosser
Master LITL, Université de Toulouse Jean Jaurès, Toulouse, France
prenom.nom@etu.univ-tlse2.fr

RÉSUMÉ

Le domaine de l'alignement sémantique automatique consiste à mettre en correspondance des entités issues de ressources d'un même domaine présentant des degrés de formalisation variés. L'alignement de ressources constitue un enjeu pour l'interopérabilité et l'exploitation de données. Les nomenclatures professionnelles sont des références pour l'organisation et la description des activités de travail pourtant, elles présentent des structures diverses et des variations lexicales importantes. Notre travail propose une méthode originale de mise en correspondance de ces nomenclatures basée sur un alignement sémantique des tâches associées aux différents métiers plutôt que sur les libellés mêmes des métiers. Pour répondre, nous avons présenté les travaux antérieurs, extrait automatiquement les tâches, puis proposé un alignement fondé sur un modèle *transformer*, aboutissant à l'alignement de trois ressources. Notre travail ouvre la voie à des recherches sur l'interopérabilité sémantique des nomenclatures professionnelles et l'analyse des proximités entre les métiers et les tâches professionnelles qui leur sont associées.

ABSTRACT

Extraction and Alignment of Task and Occupation Classifications. The field of automatic semantic alignment aims to match entities from resources within the same domain that exhibit varying levels of formalization. Resource alignment is a key challenge for interoperability and data exploitation. Occupational classifications serve as references for organizing and describing work activities, yet their structures differ significantly. Our study addresses the following question: how can we move beyond approaches based solely on the similarity of job titles to produce an alignment that faithfully reflects actual professional activities and enables the comparison of classifications developed in different institutional contexts? To answer this question, we reviewed prior work, automatically extracted tasks, and then proposed a transformer-based alignment model, resulting in the alignment of three resources. This research opens avenues for further studies on the semantic interoperability of occupational classifications and on the fine-grained analysis of similarities between occupations and the professional tasks associated with them.

MOTS-CLÉS : Alignement sémantique, Nomenclatures professionnelles
KEYWORDS : Semantic alignment, Occupational classifications

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 151–162.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1. Introduction

L'INRS (Institut National de la Recherche et de la Sécurité) est un acteur majeur de la prévention des accidents du travail et des maladies professionnelles. Parmi ses missions figure l'étude de la polyexposition et ses effets sur la santé : certains métiers cumulent en effet plusieurs facteurs de risque qu'il est crucial de prendre en compte afin de créer des campagnes de prévention complètes. Notre étude, réalisée dans le cadre d'un projet tutoré du Master Linguistique, Informatique et Technologies du Langage, est dirigée par le responsable de l'étude sur la polyexposition à l'INRS et s'inscrit dans cette démarche. Elle propose une méthodologie d'alignement de nomenclatures professionnelles fondée sur les tâches associées aux activités de travail (par exemple « plaider en justice »), plutôt que sur les libellés de métiers (« Avocats »). L'objectif est de construire une ressource unifiée et interopérable permettant de croiser plusieurs nomenclatures hétérogènes, ce qui favorise l'analyse des risques associés aux tâches exercées dans le cadre professionnel, une perspective utile pour des chercheurs en santé au travail. D'autres domaines peuvent bénéficier d'une telle ressource (éducation, sociologie)

Alors que de nombreuses approches d'alignement reposent sur la similarité lexicale ou sémantique des intitulés professionnels (Barbey et al., 2024 ; Zbib et al., 2022), nous défendons l'hypothèse qu'un alignement centré sur les tâches permet de mieux capturer la réalité des activités professionnelles et de dépasser les différences de granularité entre classifications. Pour cette étude, trois nomenclatures sont alignées : PCS (Professions et Catégories Socioprofessionnelles) produite par l'Institut national de la statistique et des études économiques (2003), ISCO (International Standard Classification of Occupation) de l'organisation internationale du Travail (2012) et le recueil des activités physiques Ainsworth (Ainsworth et al., 2011). Après avoir présenté les méthodes existantes d'alignement sémantique, nous décrivons les données utilisées, la définition opérationnelle de la notion de tâche professionnelle, puis les méthodes d'extraction de tâches au sein de la nomenclature PCS suivie de la nomenclature ISCO. Nous détaillons ensuite les alignements ISCO-PCS et ISCO-Ainsworth fondés sur des embeddings sémantiques. Enfin, nous présentons l'évaluation des extractions et des alignements, les résultats obtenus avant de conclure par un bilan et des perspectives.

2. Alignement sémantique

L'alignement sémantique de ressources vise à établir des correspondances entre des unités issues de ressources hétérogènes lorsqu'elles renvoient à des contenus similaires ou compatibles. Dans le domaine des nomenclatures professionnelles, cet alignement permet l'interopérabilité entre différentes classifications.

Dans leur article, Reimers et Gurevych (2019) présentent Sentence-BERT, une adaptation de BERT qui produit des embeddings optimisés pour la similarité sémantique de phrases et la recherche efficace de voisins proches dans de grands corpus. Cette approche nous semble pertinente pour notre travail d'alignement sémantique de nomenclatures, en utilisant un modèle multilingue.

Par ailleurs, la littérature récente explore des problématiques connexes, telles que l'articulation entre titre de métier et compétences (Zbib et al., 2022), ou l'extraction de qualifications au sein de données bruitées (Bocharova et al, 2023). Ces recherches montrent que la mise en commun de diverses sources d'information est une tâche complexe.

Une première méthode d'alignement a été appliquée à deux des nomenclatures mobilisées dans notre étude. Dans leur étude, Barbey et al. (2024) utilisent trois modèles d'embeddings pour aligner au niveau des libellés métiers les nomenclatures PCS et ISCO. Le modèle multilingue pré-entraîné bge-multilingual- gemma2 obtient les meilleures performances sur la base d'un gold de 20 items annotés par 4 experts. Les résultats montrent que 75% des alignements sont parfaits ou presque parfaits, et 20% de résultats mitigés. Le Kappa de Fleiss obtenu (0,57 et 0,52 selon le critère retenu) met en valeur la difficulté du travail d'annotation.

Notre travail s'inscrit dans cette continuité méthodologique mais se distingue par une approche fondée sur les tâches. Celle-ci propose de baser l'alignement des ressources sur les tâches décrites pour chaque métier plutôt que sur le seul libellé des professions car elles fournissent plus d'informations sémantiques. Cet angle d'analyse vise ainsi à fonder l'alignement sur une comparaison plus fine et plus opérationnelle des activités professionnelles et des risques associés.

3. Présentation des données

Ce travail mobilise des nomenclatures, c'est-à-dire des documents permettant de classer des termes (ici, des métiers) et servant de référence pour leur description et leur classification. Plus précisément, il recourt à trois nomenclatures : PCS (2003), ISCO (2008) et le Compendium of Physical Activities d'Ainsworth (2024). PCS (2003) décrit 538 métiers en français, organisés selon une structure hiérarchique. Chaque métier est associé à un code, un libellé et une description synthétique. Les descriptions présentent une longueur moyenne de 64 mots, avec des valeurs variant entre 0 et 607 mots (certaines entrées ne comportant pas de description textuelle). ISCO (2008) recense 619 professions en anglais, également structurées de manière hiérarchique. Chaque entrée comporte un code, un libellé, une description synthétique et une liste structurée de tâches. Les descriptions présentent une longueur moyenne de 37 mots, avec des valeurs comprises entre 9 et 169 mots. Compendium of Physical Activities of Ainsworth (2024) constitue une ressource en anglais recensant 1111 activités physiques associées à un libellé et à une valeur de dépense énergétique (MET : Metabolic Equivalent), qui quantifie l'intensité de l'effort physique impliqué. Les formulations associées à ces activités sont plus brèves, avec une longueur moyenne d'environ 7 mots (minimum : 1, maximum : 29).

Les exemples suivants illustrent les informations fournies par chacune de ces ressources :

1. PCS (2003) - Artisans coiffeurs (code 217c) : procédant aux soins et traitements capillaires (…)
2. ISCO (2008) - Hairdressers (code 5141) : cut, style, colour hair; shave or trim facial hair
3. Compendium of Ainsworth (2024) - Hairstyling, standing : MET = 2.5

Ces ressources hétérogènes diffèrent par leur finalité, leur langue (seule la nomenclature PCS est en français) et leur granularité, PCS et ISCO décrivent des métiers à des niveaux de détail variables, tandis qu'Ainsworth fournit une description fine d'activités élémentaires. Cette hétérogénéité constitue un enjeu central pour la constitution d'une ressource unifiée et relève du domaine de l'alignement sémantique.

4. Extraction des tâches professionnelles et alignement des nomenclatures

4.1. Identification des extractions des tâches professionnelles

Notre méthode se compose de deux étapes principales. La première consiste à identifier et extraire les tâches exprimées dans les descriptions des métiers fournies par les nomenclatures ISCO et PCS. La nomenclature Ainsworth n'est pas concernée par cette étape, car cette ressource est par nature organisée autour de la notion de tâche. La seconde étape consiste à aligner les tâches extraites en utilisant ISCO comme nomenclature pivot. Cette étape implique un alignement multilingue (anglais-français) pour ISCO-PCS et un alignement monolingue pour ISCO-Ainsworth.

4.1.1. Définition de la notion de tâche

La première étape consiste à identifier et extraire les tâches associées à chaque métier. Sur la base de définitions de dictionnaires et de l'observation d'un échantillon de tâches listées dans PCS, la définition suivante a été proposée et validée par l'expert, commanditaire de notre travail : « Action concrète effectuée dans un cadre particulier de travail, une activité professionnelle. Elle décrit une action, un travail ou une mission à exécuter qui pourrait être présentée dans une fiche de poste. Elle est caractérisée par un objectif (un objet ou résultat précis) et peut donc être composée de plusieurs sous-tâches. » Parmi des exemples de tâches professionnelles, on peut citer « négocier des contrats » (PCS, 2003) ou « the application of scientific or artistic concepts and operational methods » (ISCO, 2008). La notion de tâche professionnelle étant définie, l'extraction des tâches dans les descriptions des nomenclatures a été envisagée différemment pour les deux ressources.

4.1.2. Extraction des tâches dans la nomenclature ISCO

La nomenclature ISCO fournit pour chaque métier une liste explicite des tâches effectuées. La nomenclature et les listes de tâches associées étant rédigées par des experts, nous avons fait le choix de nous appuyer sur le découpage explicite proposé sans chercher à segmenter les tâches plus finement. Les listes de tâches suivent des structures légèrement différentes selon le niveau du métier. Pour les niveaux 1 à 3, les tâches sont introduites par une phrase et séparées par des points-virgules, comme dans l'exemple suivant : Tasks performed by professionals usually include: conducting analysis and research [...] ; advising on or applying existing knowledge [...] ; [...]. Pour les métiers de niveau 4, les tâches sont listées sous forme de puces (a), (b), (c), etc. À partir de ces listes, une extraction des tâches est réalisée à l'aide d'expressions régulières (regex) adaptées à ces deux structures, suivie d'un léger nettoyage (normalisation de la casse et de la ponctuation).

4.1.3. Annotation des tâches dans PCS

L'extraction des tâches à partir des descriptions PCS (2003) constitue une étape clé pour transformer des informations textuelles en unités exploitables. Cette phase vise à stabiliser la définition d'une tâche et à préparer l'automatisation du processus.

Une extraction manuelle des tâches PCS a été réalisée par huit annotateur·rice·s : sept étudiant·es de Master LITL ainsi qu'un expert du domaine (INRS). Un échantillon aléatoire de descriptions PCS a été sélectionné, complété par l'ajout de cas ciblés fournis par l'expert. Ces cas ont été retenus car les métiers qu'ils décrivent couvrent environ 20 % de la population française active sur la période 2003-2020 et garantissent la représentativité des situations professionnelles. Au total, 25 descriptions ont été annotées sur la plateforme Inception (Klie et al., 2018). Pour chaque description, les annotateur·rice·s devaient sélectionner les segments correspondant à la notion de tâche telle que nous l'avions définie. Aucune indication n'avait été donnée quant aux règles de délimitation. Le calcul de l'Alpha de Krippendorff indique un accord inter-annotateur compris entre 0.08 et 0.73, avec une moyenne de 0.43, ce qui témoigne d'une variabilité importante et justifie la mise en place d'une phase d'adjudication. Les principaux désaccords concernent le découpage de tâches en sous-tâches ou leur conservation en un seul segment, l'extraction de parenthèses utiles à l'explicitation d'une série de tâches et d'éléments de contexte, ainsi que la finalité des tâches exprimée dans la description. Les annotations adjudiquées ont ensuite été mobilisées pour la conception du système d'extraction automatique des tâches dans PCS. Les décisions prises lors de cette phase ont permis de clarifier les critères d'identification des tâches et de sélectionner des exemples annotés issus des descriptions adjudiquées. Ces exemples ont été intégrés au prompt utilisé pour guider le modèle dans l'identification des segments correspondant à des tâches professionnelles.

4.1.4. Extraction des tâches dans PCS

L'extraction automatique des tâches à partir des descriptions PCS repose sur l'utilisation d'un grand modèle de langue (LLM), LLaMA 3.1, mobilisé via une API. Un unique prompt a été utilisé pour cette extraction. Celui-ci s'appuie sur la définition de la notion de tâche ainsi que sur cinq exemples annotés issus de l'annotation, comprenant à la fois des exemples positifs (segments correspondant à des tâches) et négatifs (segments ne correspondant pas à des tâches). Il précise que seules les portions de texte présentes dans la description doivent être extraites. Les exemples fournis illustrent les décisions prises lors de la phase d'adjudication, notamment en ce qui concerne le découpage des tâches, l'inclusion éventuelle de parenthèses explicatives ou l'exclusion d'éléments de contexte. Pour chaque description, le modèle génère une sortie au format JSON comprenant le code PCS, le niveau hiérarchique, le texte de la description analysée, la liste des tâches identifiées ainsi que leur nombre. Les hyperparamètres ont été configurés afin de favoriser la stabilité des sorties. Un maximum de 2 000 tokens par texte analysé a été appliqué afin de capturer l'ensemble des tâches potentiellement longues dans une seule réponse. Une température de 0.2 permet d'assurer que le modèle soit le plus déterministe possible et maximise la reproductibilité des expériences.

4.1.5. Bilan de l'extraction des tâches

L'extraction des tâches a été appliquée aux nomenclatures ISCO et PCS (Table 1). Dans le cas de la nomenclature ISCO, 603 métiers ont été traités, permettant d'identifier 3 868 tâches uniques, soit une moyenne de 7,44 tâches par métier (minimum : 1 ; maximum : 16). Parmi ces tâches, 3 469 (89,68 %) sont spécifiques à un métier, tandis que 399 (10,32 %) apparaissent dans plusieurs descriptions. Ce faible taux de recouvrement doit cependant être nuancé : la nomenclature ISCO n'est pas strictement normalisée et certaines tâches diffèrent uniquement par des variations lexicales mineures. Dans la nomenclature PCS, les 538 entrées ont été traitées, parmi lesquelles 531 contenaient une description de poste. L'extraction automatique a permis d'identifier 1 924 occurrences de tâches, soit une moyenne de 3,58 tâches par description. Après

suppression des doublons, 1 681 tâches distinctes ont été obtenues. Parmi celles-ci, 1 507 tâches (89,6 %) sont spécifiques à un code PCS, tandis que 174 tâches (10,4 %) apparaissent dans plusieurs descriptions de métiers. La description comportant le plus de tâches en contient 22 (Techniciens de recherche-développement et des méthodes de production des industries de transformation, code 475a).

	ISCO	PCS	Ainsworth
Nombre de tâches	3 868	1681	1 111
Tâches uniques à un code	3 469 (89.7%)	1 507 (89.6%)	-
Tâches communes à plusieurs codes	399 (11.3%)	174 (11.4%)	-

TABLE 1 : Bilan du nombre de tâches par nomenclature

Parmi les tâches qui apparaissent dans plus de cinq métiers différents, on retrouve par exemple « assurent des fonctions d'encadrement » ou « coordination avec les autres services » dans la nomenclature PCS. Dans ISCO, des formulations similaires apparaissent, telles que « supervision of other workers » ou encore « planning and directing daily operations ». Ces tâches renvoient toutes à des activités d'encadrement et de gestion d'équipe. En se plaçant au niveau des activités réelles de travail, il devient ainsi possible de dépasser les différences de granularité entre nomenclatures, de mieux rendre compte des proximités entre métiers distincts et de proposer une approche alternative pour l'analyse des expositions professionnelles.

4.2. Alignement automatique des nomenclatures

L'alignement des nomenclatures ISCO et PCS s'appuie sur les tâches extraites précédemment. Dans ce travail, la nomenclature ISCO est utilisée comme pivot pour l'alignement des ressources. Ce choix s'explique notamment par le fait qu'elle propose, pour chaque profession, une liste explicite de tâches. Deux alignements sont réalisés indépendamment : ISCO-PCS et ISCO-Ainsworth. Dans les deux cas, les tâches extraites sont représentées sous forme vectorielle à l'aide d'un modèle Sentence-BERT (Reimers & Gurevych, 2019). Un modèle multilingue (paraphrase-multilingual-mpnet-base-v2) est utilisé pour l'alignement ISCO-PCS afin de représenter conjointement les tâches en anglais et en français, tandis qu'un modèle équivalent monolingue anglais est utilisé pour l'alignement ISCO-Ainsworth.

Chaque tâche est ainsi projetée dans un espace vectoriel commun et la similarité entre deux tâches est calculée à partir de la similarité cosinus entre leurs vecteurs. Pour chaque tâche ISCO, les tâches les plus similaires dans la nomenclature cible sont proposées comme correspondances candidates. Deux paramètres contrôlent la génération des correspondances. Le premier consiste à limiter le nombre maximal de tâches proposées pour chaque tâche ISCO à cinq. Ce seuil a été choisi après observation du nombre moyen de correspondances obtenues sans limitation, qui s'élève à environ quatre propositions par tâche. Le second paramètre correspond au seuil minimal de similarité cosinus permettant de considérer deux tâches comme similaires. Afin de déterminer une valeur appropriée pour ce seuil, nous avons analysé l'effet de différents seuils de similarité sur deux indicateurs : le nombre de correspondances générées et la couverture des tâches ISCO. La couverture correspond à la proportion de tâches ISCO pour lesquelles au moins une correspondance est identifiée, tandis que le nombre de correspondances correspond au nombre total de paires de tâches jugées similaires pour un seuil donné.

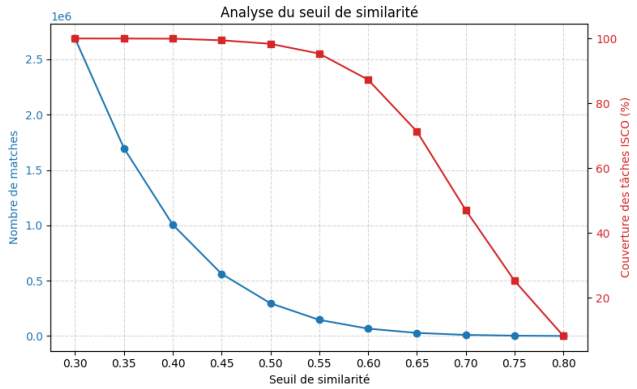

Figure 1 : Analyse du seuil de similarité pour l'alignement ISCO-PCS

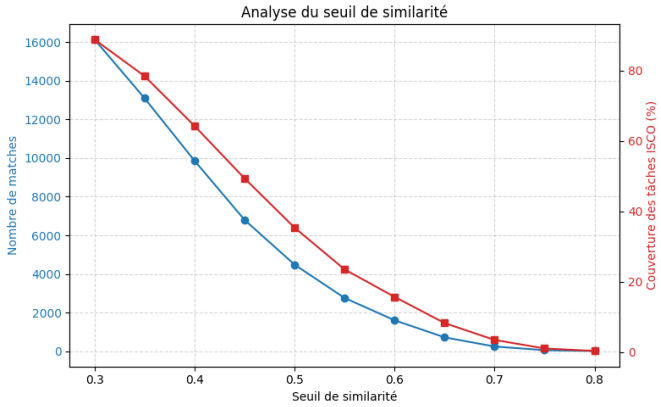

Figure 2 : Analyse du seuil de similarité pour l'alignement ISCO-Ainsworth

Les Figures 1 et 2 présentent l'évolution du nombre de correspondances et de la couverture des tâches ISCO en fonction du seuil de similarité. Les échelles diffèrent entre les deux figures en raison du nombre plus important de correspondances potentielles dans l'alignement ISCO-PCS que dans l'alignement ISCO-Ainsworth. On observe que l'augmentation du seuil de similarité entraîne une diminution rapide du nombre de correspondances ISCO-PCS ainsi que de la couverture des tâches ISCO. Le seuil de 0,5 retenu pour ce travail se situe avant la chute marquée de la couverture des tâches et permet de maintenir un compromis entre qualité des correspondances et couverture du corpus. Pour l'alignement ISCO-Ainsworth, l'application de ce seuil conduit à une couverture d'environ 35 % des tâches ISCO lorsque l'ensemble des correspondances est considéré. Cette couverture relativement limitée était attendue, la nomenclature Ainsworth étant principalement centrée sur des activités physiques spécifiques et présentant donc un recouvrement structurellement plus faible avec une nomenclature de métiers généraliste comme ISCO.

5. Résultats

5.1. Alignement ISCO-PCS

Afin d'évaluer la qualité de l'alignement ISCO et PCS, un échantillon de 87 alignements a été annoté manuellement. Chaque paire de tâches a été annotée de manière binaire : 1 si l'alignement est jugé correct, 0 sinon. L'échantillon a été constitué de manière stratifiée selon le score de similarité attribué par le modèle, avec dix alignements aléatoirement sélectionnés dans

chaque intervalle de score de 0,05. La notion de rang de la proposition a également été prise en compte dans la stratification, permettant de retrouver une répartition équitable entre les propositions 1 à 5. L'accord inter-annotateur obtenu est de 0,68 (kappa de Fleiss). Pour le cas PCS-ISCO, 21 512 alignements sont proposés. Une correspondance est proposée pour 98 % des tâches ISCO, soit 4 411 tâches, et pour 76 % des tâches PCS, soit 1 496 tâches. L'asymétrie qui apparaît ici était attendue car ISCO est la nomenclature pivot de l'alignement. Les scores varient entre les seuils de 0,5 et 0,94. Plus précisément, 689 propositions ont un score supérieur ou égal à 0,8, 15 815 ont un score supérieur ou égal à 0,6 et 5 008 ont un score inférieur à 0,6. L'évaluation d'un extrait de l'alignement des tâches ISCO-PCS permet d'observer la qualité des propositions selon le seuil (Table 2).

Seuil du score	Nombre de cas	Cas corrects	Cas corrects (%)
≥ 0.8	27	27	100
≥ 0.6	40	28	70
≤ 0.6	20	9	45
Total	87	64	74

TABLE 2 : Évaluation de l'alignement ISCO-PCS après adjudication

L'annotation des alignements par strates nous permet de remarquer que les alignements sont jugés corrects pour les scores supérieurs à 0.8 avec un taux moyen de 100% d'alignements corrects contre 70% pour les scores de 0.8 à 0.6 et 45% pour les scores inférieurs à 0.6. Les niveaux de similarité sont donc des indicateurs permettant de mieux appréhender l'alignement des ressources. Cependant, l'étude de cas particuliers révèle certaines limites. Les situations de désaccord entre les annotateur·rice·s apparaissent généralement lorsque l'élément PCS à aligner ne correspond pas à la définition de tâche. Par exemple, la tâche PCS « en vue de l'élaboration de spectacles vivants » est alignée avec la tâche ISCO « lifting and mounting scenery, lighting and other equipment in theatres and on film sets », avec un score de similarité de 0.54. De même, « en général dans le cadre d'interventions chirurgicales » (PCS), alignée à la tâche « assisting veterinarians to administer anaesthetics and oxygen during treatment » (ISCO) avec un score de similarité de 0.51. Dans ces deux cas, les tâches PCS (INSEE, 2003) ne correspondent pas clairement à la définition d'une tâche proposée précédemment, mais leur alignement semble néanmoins cohérent. Le travail étant destiné à évaluer l'alignement, nous avons choisi de conserver la note de 1 lors de l'adjudication.

Afin d'évaluer la possibilité d'une annotation automatique à grande échelle des alignements, une évaluation automatique de ces alignements a également été réalisée avec un LLM (llama3:8b) puis comparée avec les annotations manuelles. Plusieurs prompts ont été testés en faisant varier la précision de la demande et en testant l'ajout d'exemples. Le prompt qui s'est montré le plus performant est un prompt plutôt court et direct donnant un rôle au modèle et lui décrivant en une phrase son objectif. La précision obtenue est de 0.75 avec un kappa de Cohen de 0.367. L'accord reste relativement faible, le LLM ayant tendance à juger davantage un alignement comme correct plutôt qu'incorrect.

5.2. Alignement ISCO-Ainsworth

Concernant l'alignement ISCO-Ainsworth, un échantillon de 74 alignements a été annoté manuellement. Comme pour l'alignement ISCO-PCS, l'échantillon a été constitué de manière

stratifiée selon le score de similarité attribué par le modèle, avec des intervalles de 0,05. Chaque paire est annotée selon une échelle à trois niveaux : correspondance acceptable (1), partielle (0,5) ou incorrecte (0). Cette échelle diffère légèrement de celle ISCO-PCS, sur la base de l'adéquation du domaine d'activité et de la comparabilité du type d'effort physique impliqué. L'alignement ISCO-Ainsworth a été appliqué aux tâches ISCO correspondant aux métiers de niveaux 3 et 4 (4 091 entrées, soit toutes les tâches extraites dans la nomenclature ISCO) à partir des 1 111 activités de la nomenclature Ainsworth. En retenant un seuil de similarité de 0,5 et un maximum de cinq correspondances par tâche ISCO, 4 485 alignements ont été produits, pour une couverture d'environ 35.4 % des tâches ISCO. La majorité des correspondances relèvent d'un niveau de confiance faible, ce qui reflète la difficulté structurelle de l'alignement entre une nomenclature de métiers généraliste (ISCO) et une nomenclature d'activités physiques spécialisées (Ainsworth). Les correspondances correctes et partielles sont regroupées dans la Table 3 afin de calculer la proportion d'alignements acceptables.

Seuil du score	Nombre de cas	Correct	Partiel	Acceptables (%)
≥ 0.8	14	12	2	100
≥ 0.6	40	28	6	85
≤ 0.6	20	6	7	65
Total	74	46	15	82

TABLE 3 : Évaluation de l'alignement ISCO-Ainsworth après adjudication

Comme pour l'alignement PCS-ISCO, l'analyse qualitative montre également une corrélation nette entre le score de similarité et la qualité sémantique des alignements : les correspondances à forte similarité ($\geq$ 0,80) sont majoritairement jugées correctes, tandis que la proportion de faux positifs augmente sensiblement pour les scores faibles (0,50-0,60). Les scores intermédiaires correspondent plus fréquemment à des alignements partiels ou discutables. À titre d'illustration, « making beds, cleaning bathrooms » (ISCO) est correctement aligné avec « Chambermaid, hotel housekeeper, making bed, cleaning bathroom » (Ainsworth), de même que « digging and shovelling to clear ditches » avec « Shoveling, digging ditches », pour lesquels la valeur MET apparaît transférable, l'alignement portant sur des tâches du même domaine d'activité, associées à des actions de nature comparable et donc à un niveau de dépense énergétique similaire. L'évaluation manuelle de ces alignements présente par ailleurs un bon niveau d'accord inter-annotateurs ($\kappa \approx 0{,}83$), les rares désaccords ayant été facilement résolus lors de l'adjudication. Ces résultats confirment l'intérêt du score de similarité comme indicateur de qualité, tout en soulignant qu'il ne garantit pas à lui seul la validité sémantique des correspondances.

Une annotation automatique par modèle de langage (Qwen3:8b) a été comparée à l'annotation humaine dans une perspective exploratoire, afin d'évaluer la possibilité d'une annotation automatisée à grande échelle des alignements. Le modèle a été guidé par le même protocole d'annotation que celui fourni aux annotateurs humains, incluant la définition des catégories et des exemples illustratifs. Le modèle atteint un taux d'accord brut d'environ 70 % avec la version adjudiquée des annotations humaines et obtient un coefficient d'accord α de Krippendorff de 0,654, ce qui suggère un accord global modéré à bon, tout en indiquant que l'annotation automatique reste à manier avec prudence en raison d'erreurs non négligeables.

6. Bilan

Les résultats obtenus montrent que l'alignement automatique des tâches permet d'identifier un nombre important de correspondances entre les nomenclatures étudiées. Dans le cas de l'alignement ISCO-PCS, la forte couverture des tâches ISCO ainsi que les résultats de l'annotation manuelle indiquent que les scores de similarité constituent un indicateur pertinent de la qualité des correspondances proposées. Les alignements présentant un score élevé sont majoritairement jugés corrects, tandis que la proportion d'erreurs augmente lorsque le score de similarité diminue. L'alignement ISCO-Ainsworth présente des résultats plus contrastés. La couverture plus faible observée dans ce cas s'explique en grande partie par les différences structurelles entre les deux ressources. La nomenclature Ainsworth décrit en effet des activités physiques associées à des valeurs de dépense énergétique, tandis qu'ISCO décrit des professions et les tâches qui leur sont associées. Cette différence de granularité limite mécaniquement le nombre de correspondances possibles. Ces résultats mettent également en évidence certaines limites de l'approche. Les erreurs d'alignement apparaissent notamment lorsque les tâches extraites contiennent des éléments de contexte ou des formulations trop générales. Par ailleurs, l'hétérogénéité lexicale observée dans les nomenclatures complique la mesure automatique de similarité entre tâches pourtant proches du point de vue fonctionnel. Malgré ces limites, l'approche par tâches apparaît comme une stratégie pertinente pour l'alignement de ressources décrivant les activités professionnelles. En se plaçant au niveau des activités de travail plutôt qu'au niveau des libellés de métiers, il devient possible d'identifier des proximités entre professions appartenant à des catégories différentes et de rendre compte de la diversité des situations professionnelles.

7. Conclusion et perspectives

Dans ce travail, nous avons réalisé un alignement sémantique de ressources professionnelles. Notre contribution réside dans la proposition d'une approche centrée sur la tâche. À notre connaissance, cette approche n'avait pas encore été explorée. Après avoir extrait les tâches ISCO grâce à un système à base de règles puis les tâches PCS à l'aide d'un modèle de langue, nous avons proposé un alignement ISCO-PCS puis ISCO-Ainsworth basé sur un modèle transformer de type embeddings S-BERT. Cet alignement permet ainsi de relier conjointement trois ressources décrivant les activités professionnelles.

Ce travail met également en évidence la complexité de la notion de tâche professionnelle. Les phases d'extraction ont montré que les tâches peuvent correspondre à des réalités hétérogènes et sont souvent imbriquées avec des éléments de contexte. Cette observation souligne l'importance d'une clarification ontologique préalable dans les démarches d'alignement sémantique. En se plaçant au niveau des activités réelles de travail plutôt qu'au niveau des libellés de métiers, l'approche par tâches permet cependant de dépasser certaines différences de granularité entre nomenclatures et de révéler des proximités fonctionnelles entre métiers distincts.

Plusieurs pistes de prolongement peuvent être envisagées. Tout d'abord, une nouvelle campagne d'extraction manuelle pourrait être menée à l'aide d'un guide d'annotation élaboré à partir des discussions d'adjudication menées dans ce travail. Un tel guide permettrait d'obtenir une extraction plus homogène et de réduire le travail d'adjudication, tout en fournissant éventuellement une base d'apprentissage pour améliorer le prompt utilisé lors de l'extraction automatique. Par ailleurs, dans la mesure où les valeurs MET d'Ainsworth attribuées aux tâches ISCO couvrent principalement des activités physiques, il serait pertinent de mener une analyse complémentaire des métiers caractérisés par des activités plus statiques. Une telle analyse pourrait nécessiter

l'intégration d'autres ressources décrivant les activités professionnelles, par exemple Wikipédia qui propose une catégorisation des métiers par secteur et décrit parfois les tâches associées. Enfin, un alignement intra-nomenclature des tâches pourrait constituer une piste de prolongement intéressante. Bien que la nomenclature ISCO fournisse des tâches explicites, leur formulation varie au sein même de la ressource. Une normalisation ou un regroupement des formulations proches permettrait d'identifier plus clairement les tâches équivalentes ou similaires et d'améliorer la cohérence globale de l'alignement.

Notre travail ouvre ainsi la voie à de futurs travaux sur l'interopérabilité sémantique des nomenclatures professionnelles et sur l'analyse fine des proximités entre métiers.

Références

AINSWORTH B. E., HASKELL W. L., HERRMANN S. D., MECKES N., BASSETT D. R., TUDOR-LOCKE C., GREER J. L., VEZINA J., WHITT-GLOVER M. C. & LEON A. S. (2011). *2011 compendium of physical activities: A second update of codes and MET values*. *Medicine & Science in Sports & Exercise, 43*(8), 1575–1581. DOI : 10.1249/MSS.0b013e31821ece12.

BARBEY C., CLERC F., SANTOS SOUSA G., TANGUY L. & TROJAHN C. (2024). *Harmonisation des nomenclatures métier*. Rapport technique interne.

BOCHAROVA M., MALAKHOV E. & MEZHUYEV V. (2023). VacancySBERT: The approach for representation of titles and skills for semantic similarity search in the recruitment domain. arXiv. DOI : 10.48550/arXiv.2307.16638.

CATÉGORIE : MÉTIER PAR SECTEUR. (2022). Dans *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=Cat%C3%A9gorie:M%C3%A9tier_par_secteur&oldid=197014986.

HERRMANN S. D., WILLIS E. A., AINSWORTH B. E., BARREIRA T. V., HASTERT M., KRACHT C. L., SCHUNA J. M., CAI Z., QUAN M., TUDOR-LOCKE C., WHITT-GLOVER M. C. & JACOBS D. R. (2024). Adult compendium of physical activities: A third update of the energy costs of human activities. *Journal of Sport and Health Science, 13*(1), 6–12. DOI : 10.1016/j.jshs.2023.10.010.

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. (2003). *Nomenclature des professions et catégories socioprofessionnelles (PCS 2003)*. INSEE. https://www.insee.fr/fr/information/2497952.

INTERNATIONAL LABOUR ORGANIZATION. (2012). *International standard classification of occupations: ISCO-08. Volume I: Structure, group definitions and correspondence tables*. International Labour Office.

KLIE J.-C., BUGERT M., BOULLOSA B., DE CASTILHO R. E. & GUREVYCH I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, p. 5–9. https://aclanthology.org/C18-2002/.

ORGANISATION INTERNATIONALE DU TRAVAIL. (2012). *Classification internationale type des professions (CITP-08): Structure, définitions des groupes et tables de correspondance*. OIT.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), p. 3982–3992. DOI : 10.18653/v1/D19-1410.

ZBIB R., LACASA L. A., RETYK F., POVES R., AIZPURU J., FABREGAT H., SIMKUS V. & GARCÍA-CASADEMONT E. (2022). Learning job titles similarity from noisy skill labels. arXiv. DOI : 10.48550/arXiv.2207.00494.