Université Paris Cité
Master *Mathématiques et applications - Parcours Logos*
Laboratoire d'accueil : LIP6 (Sorbonne Université, CNRS)

Construction d'un graphe de connaissances et quantification de l'incertitude dans le corpus *Studium Parisiense*

Rapport de stage présenté par Narada Maugin

Stage effectué du 9 février au 9 août 2026

Septembre 2026

Camelia Constantin	Encadrante (LIP6)
Raphaël Fournier-S'niehotta	Encadrant (LIP6 & CNAM)
Cédric du Mouza	Encadrant (CNAM)

« Toute connaissance humaine est incertaine, inexacte et partielle. »

— Bertrand Russell, *Human Knowledge: Its Scope and Limits*, 1948.

Remerciements

En décrivant la fabrication d'un habit de laine, Adam Smith faisait apparaître la multitude de personnes cachées derrière un objet ordinaire. Ce rapport n'échappe pas à la règle. Je voudrais remercier ici les personnes qui m'ont accompagné pendant ce stage, ainsi que celles qui ont constitué les données sur lesquelles j'ai travaillé.

Je remercie d'abord mes encadrants, Camelia Constantin, Raphaël Fournier-S'niehotta et Cédric du Mouza. Leurs conseils et nos discussions m'ont aidé à mettre mes résultats à l'épreuve, à formuler de nouvelles questions et à prendre du recul sur mon travail.

Je remercie ensuite les personnes qui ont contribué à la rédaction et à l'enrichissement des fiches de *Studium Parisiense*, en premier lieu Jean-Philippe Genet et Thierry Kouamé. Leur travail sur les sources a été le point de départ du mien. Merci également à Stéphane Lamassé pour nos échanges. Ma reconnaissance va aussi à celles et ceux dont j'ai utilisé les contributions sans avoir eu l'occasion de les rencontrer.

Je remercie Hubert Naacke, qui m'a accueilli dans l'équipe BD, ainsi que les membres et le personnel du LIP6 pour leur bienveillance au quotidien et pour nos échanges pendant ces quelques mois. Merci également à Allaa, Anis et Nour, doctorants, ainsi qu'à Defne, Michele et Wafaa, stagiaires, avec qui j'ai partagé le bureau.

Mes remerciements vont enfin aux équipes administratives qui ont contribué au bon déroulement de ce stage. Je remercie en particulier Marie Véronique Lam, qui a préparé et organisé mon séjour au huitième symposium du GdR MaDICS à Avignon.

Résumé

Ce stage porte sur *Studium Parisiense*, une base prosopographique qui rassemble plus de 20 000 notices consacrées aux membres des écoles et de l'université de Paris du XII^e^ au XVI^e^ siècle. Il vise à convertir ces notices en un graphe de connaissances, dans lequel personnes, lieux, institutions et événements sont représentés par des nœuds et sont reliés entre eux. Il étudie également comment repérer et quantifier l'incertitude exprimée par les rédacteurs. Établies à partir de sources médiévales lacunaires, les notices mêlent en effet faits assurés, hypothèses et valeurs approximatives. Préserver ces distinctions lors de la conversion permet de conserver les réserves formulées.

Un programme de conversion fondé sur des règles a d'abord été développé pour transformer les champs structurés de l'export en un graphe Neo4j suivant le modèle prosopographique DAPHNÉ. Il a été appliqué à un sous-échantillon du corpus. Le graphe obtenu est un premier état qui possède des limites et qui a vocation à être enrichi. Cinq modèles de langage ont ensuite été comparés pour attribuer un ou plusieurs types (diocèse, collège, ville…) aux libellés du champ des institutions. Le meilleur atteint un F1 moyen par entité de 87,6 % sur les 139 libellés du test. Ces prédictions pourraient servir de propositions à valider par une relecture humaine experte, limitée aux cas où les modèles ne s'accordent pas entre eux.

Un deuxième volet prolonge un article en préparation pour la revue *Data & Knowledge Engineering*. Il compare des méthodes qui attribuent à chaque phrase un score de 1 à 5 selon un barème. Après nettoyage et réannotation du corpus, quatre méthodes de référence ont été comparées à neuf modèles de langage utilisés sans entraînement supplémentaire sur ces données. Sur les 487 phrases du test, huit de ces modèles obtiennent une exactitude supérieure à celle de CamemBERT. Les meilleurs obtiennent également un F1 macro plus élevé, mesure qui accorde le même poids aux cinq niveaux.

Le dernier volet distingue trois phénomènes : l'incertitude sur la validité d'un fait, l'imprécision d'une valeur (« environ 1248 ») et l'incomplétude de l'information (la valeur manque). L'annotation précise également leur portée : l'énoncé entier, une relation ou une entité. Son échelle de certitude est distincte du barème précédent. Sur un jeu de référence de 206 extraits de notices, les sept modèles de langage évalués obtiennent entre 88,2 et 93,5 % de F1 strict. Cette mesure exige d'identifier à la fois le phénomène et le type d'information visé. Ce volet a été présenté sous forme de poster au Symposium MaDICS 2026.

Ces résultats décrivent la conformité des prédictions au barème retenu, qui reste une convention, non la vérité historique des énoncés. Les annotations d'incertitude et les scores de certitude issus des deux derniers volets restent à verser dans le graphe, dont les arêtes offrent déjà les attributs pour les recevoir.

Table des matières

Table des figures

Liste des tableaux

Introduction

Cadre du stage

Ce stage s'est déroulé du 9 février au 9 août 2026 au LIP6, le Laboratoire d'informatique de Sorbonne Université sous la double tutelle de Sorbonne Université et du CNRS, sur le campus Pierre-et-Marie-Curie à Paris[1]. Avec environ 500 membres, dont plus de 150 enseignants-chercheurs, le LIP6 est l'un des plus grands laboratoires d'informatique de France. Ses 18 équipes de recherche sont réparties en quatre axes thématiques :

— intelligence artificielle et sciences des données
— architecture, systèmes et réseaux
— sécurité, sûreté et fiabilité
— théorie et outils mathématiques pour l'informatique

J'ai été accueilli dans l'équipe BD (Bases de données), dirigée par Hubert Naacke et rattachée aux axes « Intelligence artificielle et sciences des données » et « Architecture, systèmes et réseaux ». Elle étudie la gestion, la transformation et l'analyse des données, aussi bien par les approches classiques des bases de données que par des méthodes statistiques. Les graphes de connaissances et la qualité des données, dont relève ce stage, comptent parmi ses thèmes.

Le stage a été co-encadré par trois enseignants-chercheurs : Camelia Constantin, membre de l'équipe BD du LIP6, Raphaël Fournier-S'niehotta, responsable de l'équipe ComplexNetworks du LIP6, et Cédric du Mouza, professeur au Conservatoire national des arts et métiers (CNAM), au sein de l'équipe ISID (Ingénierie des systèmes d'information et de décision) du laboratoire CEDRIC.

Le stage s'inscrit dans un axe de recherche sur les bases prosopographiques coordonné par Cédric du Mouza. Le projet ANR DAPHNÉ[2] a porté sur la modélisation et l'exploitation de ces corpus. Le projet ANR LAURA[3], lancé en décembre 2025, le prolonge en étudiant l'enrichissement d'une base de connaissances à partir de données incertaines, et le stage s'y rattache.

1. https://www.lip6.fr/ (consulté le 01/09/2026)
2. « Découverte dans les bAses Prosopographiques Historiques de coNnaissancEs », ANR-17-CE38-0013 : https://anr.fr/Projet-ANR-17-CE38-0013 (consulté le 01/09/2026).
3. « Enrichissement d'une base de connaissances à partir de données prosopographiques médiévales incertaines », ANR-25-CE38-0700 : https://anr.fr/Projet-ANR-25-CE38-0700 (consulté le 01/09/2026).

Problématique et objectifs

La prosopographie étudie un milieu social à travers l'ensemble de ses membres, et non à travers quelques figures illustres. Les bases prosopographiques regroupent ces informations dans des notices que l'historien peut interroger pour compter, comparer et rechercher des régularités [3]. On peut ainsi examiner si les maîtres d'une discipline, comme les arts, la médecine ou le droit canon, viennent de régions particulières [41].

Studium Parisiense[4] en est une, consacrée aux membres de l'université de Paris au Moyen Âge. Chacune de ses plus de 20 000 notices rassemble ce que des sources médiévales dispersées et lacunaires laissent savoir d'un individu.

Ces notices sont rédigées en texte libre par des historiens. La formulation renseigne sur le statut de l'information. « Il naît à Meulan », « il naît peut-être à Meulan », « selon Duchesne il naît à Meulan » et « il naît vers 1248 à Meulan » ne s'affirment pas au même titre : la première énonce, la deuxième suppose, la troisième précise l'origine de l'affirmation, la quatrième approxime. Les verser toutes les quatre dans un graphe de connaissances sous une forme identique reviendrait à effacer ce que l'historien a pris soin d'écrire, et à faire passer pour acquis ce qui n'était qu'une conjecture.

Le chapitre 1 présente la base et le modèle DAPHNÉ, et le chapitre 2 situe le stage dans la littérature. Le chapitre 3 présente la conversion des champs structurés de la base en une première version du graphe de connaissances. Il étudie aussi le typage du champ des institutions (université, collège, diocèse, ordre religieux…). Ce chantier a rendu le problème de l'incertitude inévitable, car dans les champs structurés les dates et les lieux sont souvent approximatifs, et dans les champs libres les phrases sont souvent nuancées.

Traiter cette incertitude a occupé le reste du stage, en deux volets. Le premier, au chapitre 4, attribue à chaque phrase un score de certitude de 1 à 5 selon un barème. Il prolonge un travail engagé avant le stage pour un article en préparation dans la revue *Data & Knowledge Engineering* (DKE), où quatre méthodes sont comparées sur un jeu de phrases annotées à la main, qui tient lieu de référence pour l'évaluation. Le second, au chapitre 5, prend un élément d'une fiche, détecte l'incertitude, le délimite et cherche à savoir sur quoi le doute porte, l'énoncé entier, une relation précise ou une entité. La certitude y constitue l'un des trois axes étudiés, avec la précision et la complétude. Le chapitre 6 discute ce que ces résultats montrent et impliquent, et le chapitre 7 dresse le bilan du stage et ses perspectives.

4. http://studium-parisiense.univ-paris1.fr/ (consulté le 01/09/2026). La base est présentée plus en détail au chapitre 1.

Première partie

Contexte et état de l'art

Chapitre 1

Le corpus Studium Parisiense et le modèle DAPHNÉ

Il est possible qu'il s'agisse de deux personnes distinctes.

Studium Parisiense, fiche n° 22055, Galterius de Sancto Quintino 1

1.1 La base Studium Parisiense

Studium Parisiense est développée collectivement au LAMOP (Laboratoire de médiévistique occidentale de Paris) à l'Université Paris 1 Panthéon-Sorbonne [26, 24]. Son périmètre couvre les membres des écoles et de l'université de Paris, du XII^{e} au XVI^{e} siècle. Il comprend les professeurs et les étudiants, ainsi que d'autres personnes au service de l'institution. En mai 2026, la base proposait 20 307 fiches à la consultation et l'équipe du projet en attend à terme plus de 40 000 [25].

Chaque fiche décrit un individu à travers des rubriques normalisées : fiche d'identité (nom et variantes, dates de vie et d'activité, statut), origine et situation géographique, insertion relationnelle (famille, amis, ennemis), cursus, carrières ecclésiastique et professionnelle, voyages, production textuelle (œuvres avec manuscrits, incipits et éditions), et bibliographie. Le texte libre de ces rubriques conserve les réserves du rédacteur, par exemple « peut-être frère du couvent franciscain d'Uppsala », ainsi que les valeurs approximatives, comme « vers 1248 ». Des conventions de saisie facilitent aussi le repérage des informations : les noms de personnes sont encadrés de $, les lieux sont marqués d'un * et les dates sont entre %. La notation `:1373:` signifie « vers 1373 ». Selon le contexte, le « ? » signale un doute ou une information manquante. Cette incertitude étant constitutive de la recherche historique, elle n'est pas un défaut spécifique à cette base, si tant est que l'absence de certitude en soit un. C'est une information historiographique en soi, qu'il faut savoir re-

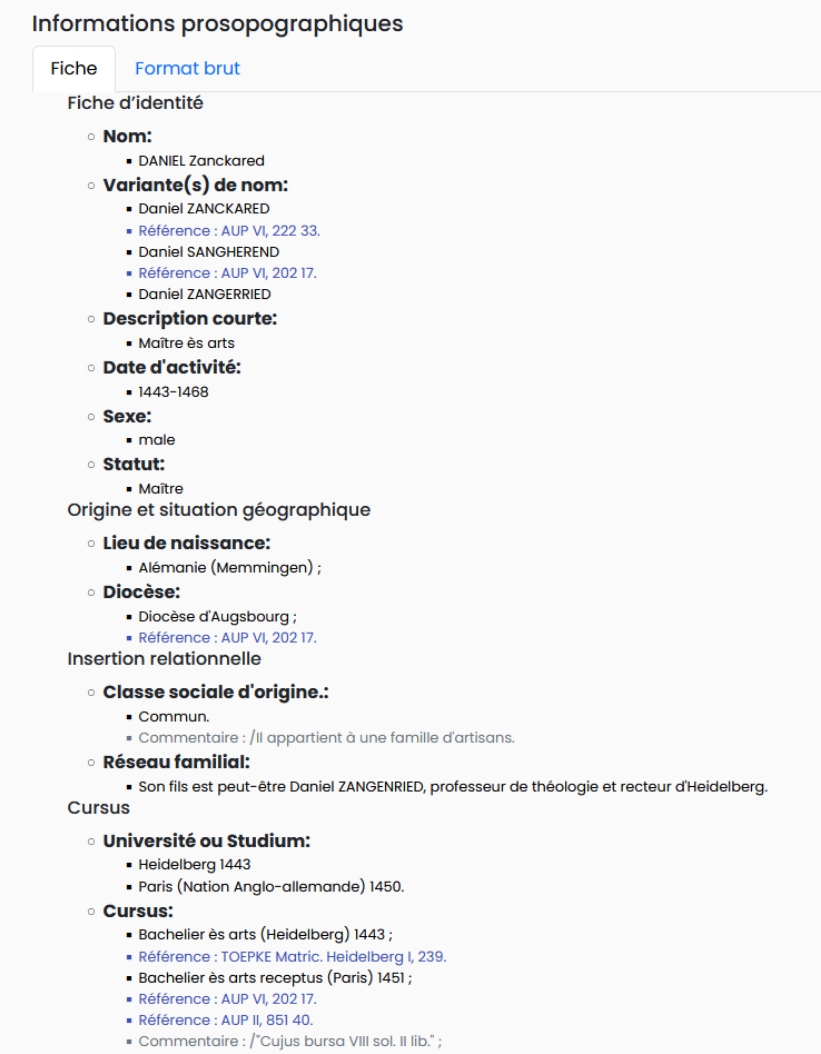

Figure 1.1 – Le début de la fiche de Daniel Zanckared. L'incertitude y est visible, « Son fils est *peut-être* Daniel ZANGENRIED », de même que les commentaires et les références qui accompagnent chaque rubrique.

pérer et quantifier [3]. La figure 1.1 montre le début d'une fiche[1], telle que le site public l'affiche, où de l'incertitude est visible au niveau du réseau familial.

1.1.1 Anatomie d'une fiche

Le stage s'est appuyé sur un export de la base au format JSONL, un fichier texte dont chaque ligne contient une fiche complète. Les rubriques ne sont pas toutes renseignées pour tout le monde. Environ une fiche sur deux, par exemple, ne dit rien de l'origine géographique de l'individu.

1. http://studium-parisiense.univ-paris1.fr/individus/1862-danielzanckared (consulté le 01/09/2026)

Sous cette présentation lisible, chaque champ répète une même structure, que l'extrait de la figure 1.2 montre sur une fiche réelle. Le volet `value` porte la phrase de l'historien, en texte libre. Le volet `meta` regroupe des noms, des lieux, des institutions et des dates déjà représentés sous forme structurée. Les dates sont dotées d'un type : exacte, intervalle, avant, après ou environ. Les champs facultatifs `comment` et `reference` contiennent respectivement les remarques de l'historien et les références associées à l'information. Ces volets viennent tels quels de la saisie dans Studium, sans transformation à ce stade. L'export entier compte 440 000 champs `value` et 156 000 dates extraites.

```
{ "identity": {
    "name":   [ { "value": "EUSTASIUS de Mesy", "meta": {} } ],
    "status": [ { "value": "Incertain",         "meta": {} } ],
    "datesOfActivity": [ { "value": "%1263-1272%",
      "meta": { "dates": [ { "type": "INTERVAL",
        "startDate": { "type": "SIMPLE", "date": 1263 },
        "endDate":   { "type": "SIMPLE", "date": 1272 } } ] } } ] },
  "relationalInsertion": {
    "socialClassOrigin": [ { "value": "Noble ;", "meta": {} } ],
    "familyNetwork": [ {
      "value": "Ses parents sont les vicomtes de Meulan ;",
      "meta": { "institutions": [ "Meulan" ] } } ],
    "personalServicesRelationship": [ {
      "value": "Service d'$Alphonse de POITIERS$ %1263-1268% ;",
      "meta": { "names": [ "Alphonse de POITIERS" ],
                "dates": [ ... ] },
      "reference": [ "FEG: II, Rouen n° 4193 (V. TABBAGH)." ] } ] },
  ... }
```

Figure 1.2 – Extrait abrégé d'une fiche de l'export JSONL. Chaque champ combine le texte libre de l'historien (`value`) et les entités qui en ont été extraites (`meta`).

1.1.2 Ce que cette structure implique pour le graphe

L'extrait de la figure 1.2 montre ce que la structure permet de convertir directement et ce qui demande une analyse supplémentaire. Dans `datesOfActivity`, `meta.dates` fournit le type de date et les bornes de l'intervalle. La conversion en graphe est simple, puisque ces valeurs peuvent être traitées par des règles explicites.

D'autres informations ne sont que partiellement structurées. Dans `familyNetwork`, le texte indique que les parents de l'individu sont les vicomtes de Meulan, mais `meta` ne contient que la mention « Meulan », rangée parmi les institutions. Le lien familial et le titre de vicomte ne sont pas représentés séparément. Leur extraction demanderait donc d'analyser le texte libre.

À l'échelle de l'export, `meta.institutions` mélange ainsi, sans les distinguer, nations universitaires, disciplines, diocèses, collèges, abbayes, ordres et villes. Les nations étaient les quatre groupements, France, Picardie, Normandie et Angleterre (devenue la na-

tion anglo-allemande), entre lesquels l'université répartissait maîtres et étudiants selon leur origine géographique. On y trouve 5 635 valeurs distinctes une fois les balises de saisie retirées, dont 3 637 n'apparaissent qu'une seule fois. Redonner un type à ces valeurs est l'objet de la section 3.5.

Les métadonnées ne conservent toutefois pas toutes les réserves du texte. Certains types de dates rendent déjà une approximation explicite, mais des formulations comme « peut-être », les « ? » et les dates en « vers » ne laissent aucune trace dans `meta`. Ces formulations demandent une analyse supplémentaire, au risque d'affirmer tout au même titre, l'attesté comme le conjecturé.

La base contient aussi des fiches qui portent le même nom, 47 cas sur l'export complet et ces cas ne sont pas de même nature. Pour 23 d'entre eux, les fiches affichent les mêmes dates d'activité et les mêmes rubriques sous deux adresses distinctes du site. Ce sont des doublons. Dans les 24 autres cas, les dates diffèrent. Quatre fiches se nomment « Henricus de Colonia » et décrivent quatre bacheliers ès arts actifs en 1339, 1349, 1372 et 1395. Six « Guillelmus Mestre d'école » se partagent de même les années 1292 à 1313. Le nom seul ne permet donc pas de décider si deux fiches parlent de la même personne. Au-delà des cas où le nom coïncide, le rapprochement des fiches qui pourraient décrire un même individu sous des noms différents reste une perspective ouverte du projet.

S'y ajoutent les accidents ordinaires d'une base saisie collectivement sur de longues années. L'export contient des lignes au JSON invalide, des intervalles de dates inversés et quelques artefacts d'encodage.

1.2 Le modèle DAPHNÉ et la représentation par factoïdes

Le projet ANR DAPHNÉ a proposé le modèle cible de cette conversion, un modèle conceptuel générique pensé pour n'importe quelle base prosopographique et non pour le seul Studium [3]. Ses auteurs l'ont éprouvé en y transposant trois bases bâties sur des principes différents : PASE [2], qui recense les habitants de l'Angleterre documentés par les sources, de la fin du VI^{e} à la fin du XI^{e} siècle, PADU-A [3], qui couvre les étudiants et les enseignants des deux premiers siècles de l'université de Padoue, de 1222 à 1405, et Studium Parisiense. La figure 1.3 en donne le schéma dans la version utilisée pendant le stage, soit 25 entités et 33 relations.

Sa particularité tient à l'entité placée au centre du schéma, le *factoïde*. Un factoïde relie une source, un fait et une personne, la source S attestant que le fait F peut être énoncé à propos de l'entité E. Il permet ainsi de distinguer une affirmation de sa validation historique. Le modèle n'enregistre donc pas « X est né à Y », mais « telle source affirme que X est né à Y ». Deux sources qui se contredisent produisent deux factoïdes, tous deux conservés, chacun rattaché à son auteur. Chaque naissance, chaque grade, c'est-à-dire chaque titre

2. https://pase.ac.uk/ (consulté le 01/09/2026)
3. https://www.dissgea.unipd.it/padu-prosopographical-access-database-university-agenda-verso-una-banca-dati-di-studenti-e-docenti (consulté le 01/09/2026)

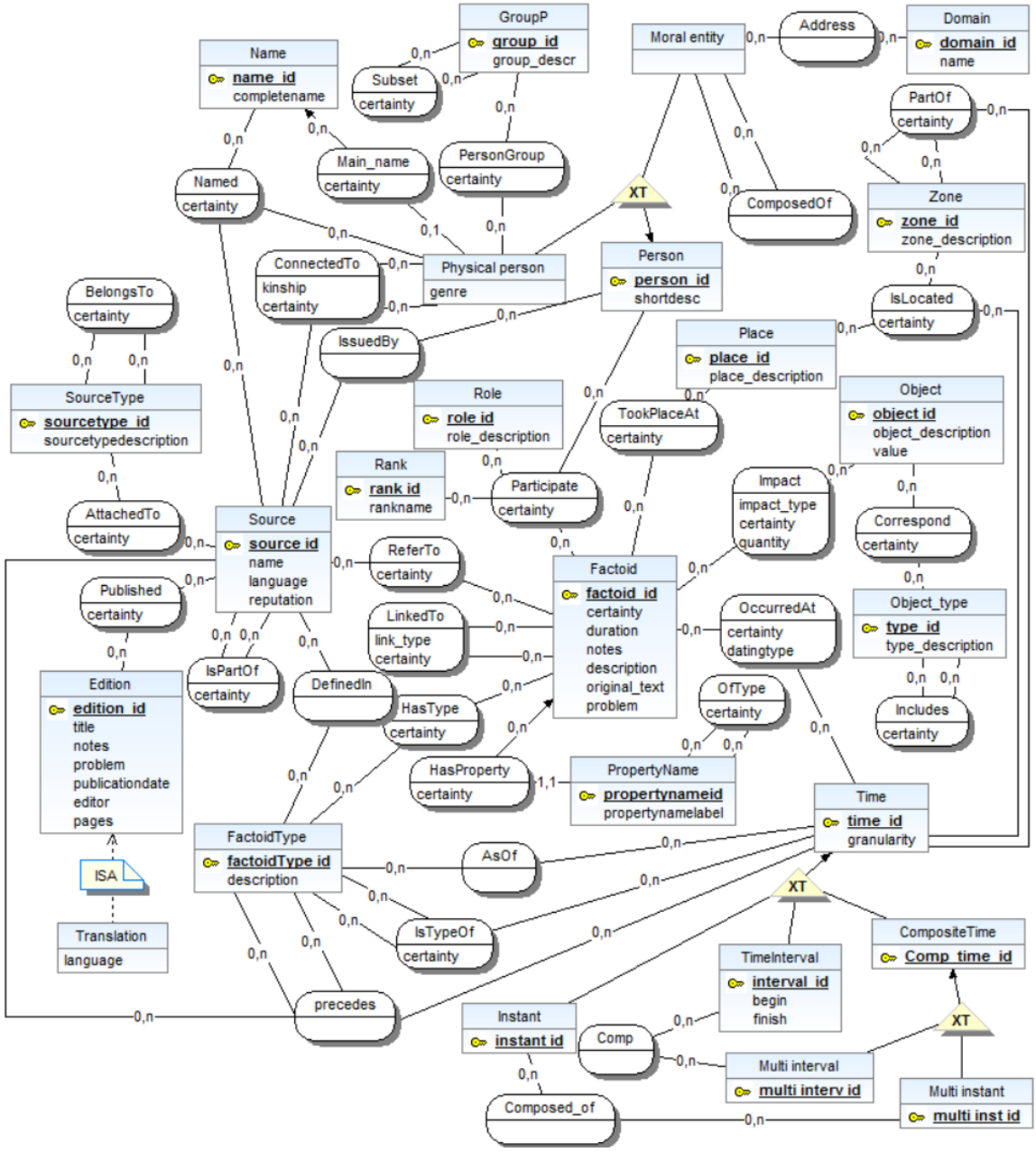

Figure 1.3 – Le schéma conceptuel du modèle DAPHNÉ (25 entités, 33 relations) issu de [3]. Les rectangles sont les entités, les ovales les relations, les triangles XT des spécialisations exclusives. L'entité Factoid, au centre, relie une affirmation à la personne concernée, au lieu, au temps et à la source qui l'atteste. L'attribut certainty revient sur la majorité des relations.

universitaire obtenu, chaque œuvre devient ainsi un nœud à part entière, relié à la personne, au lieu, au moment et à la source, et le modèle prévoit un attribut `original_text` pour la phrase dont le factoïde est issu.

La fiche d'Andreas Freouati[4] contient, dans le champ `datesOfActivity`, la valeur « %1329-1348% », et son volet `reference` indique d'où l'historien la tient, le *Chartularium Universitatis Parisiensis*, tome II, acte n° 1184. La figure 1.4 représente cette information par un factoïde relié à la personne concernée, au type « période d'activité », à l'intervalle 1329-1348 et à la source citée. Ce lien conserve la provenance associée à l'information dans la notice.

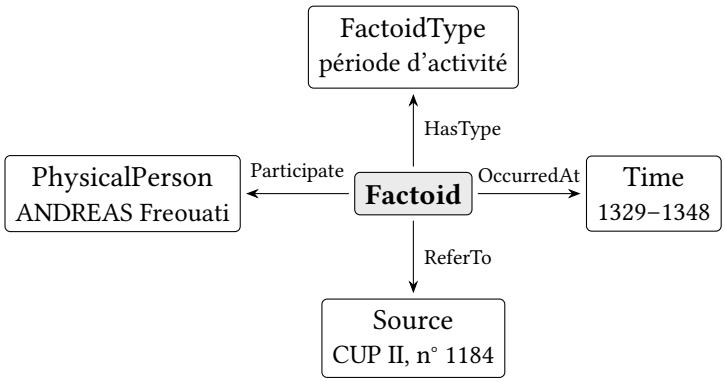

Figure 1.4 – La ligne `datesOfActivity` de la fiche d'Andreas Freouati telle que le modèle la représente.

Le type du factoïde précise la nature de l'information. Une personne peut être reliée à Paris pour sa naissance, l'obtention d'un grade ou un enseignement. Distinguer ces types permet de formuler des requêtes ciblées. Pour rechercher les personnes ayant obtenu un grade à Paris avant 1300, il faut sélectionner les factoïdes de grade, puis leurs liens avec le lieu et la date.

La distinction entre le factoïde et les entités concernées permet aussi de préciser la portée du doute. Dans la figure 1.1, « Son fils est peut-être Daniel ZANGENRIED » met en doute la filiation proposée. L'annotation doit donc viser l'information de filiation. DAPHNÉ prévoit un attribut `certainty` sur le factoïde et sur 25 de ses 33 relations. La relation temporelle `OccurredAt` possède également un attribut `datingtype`, qui distingue notamment une date exacte, une date approximative et une date connue par une borne. Ces catégories permettent de représenter les types de dates déjà présents dans `meta.dates`.

Le modèle DAPHNÉ fournit ainsi des attributs pour représenter la certitude et la précision des informations. Il reste à déterminer comment les renseigner à partir des notices. Certaines indications peuvent être reprises des champs structurés, d'autres nécessitent d'interpréter un marqueur et sa portée dans le texte libre.

4. http://studium-parisiense.univ-paris1.fr/individus/640-andreasfreouati (consulté le 01/09/2026)

Chapitre 2

État de l'art

L'attribution à Robert Kilwardby est douteuse ; peut-être est-ce une oeuvre de Jean de Sècheville ?

Studium Parisiense, fiche n° 51775, Robertus Kilwardby

Ce chapitre situe les choix du stage dans les travaux existants. Les sections 2.1 et 2.2 présentent les modèles de bases prosopographiques et les graphes de connaissances. Les sections 2.3 à 2.7 examinent les formes de l'incertitude, leur représentation dans les données et leur détection dans les textes. La section 2.8 introduit l'apprentissage en contexte des modèles de langage, utilisé dans les expériences du rapport. Enfin, la section 2.9 présente le rapprochement d'entités, envisagé comme une suite du travail, avant le bilan du chapitre.

2.1 Les bases prosopographiques et leurs modélisations

Il existe plusieurs manières de modéliser une base prosopographique [2]. Dans une base *relationnelle*, l'information est rangée dans des tables, comportant une ligne par personne ou par événement et une colonne par attribut. Chaque ligne porte un numéro qui l'identifie, et les tables se relient entre elles par ce numéro. Une table des œuvres indique, pour chaque titre, le numéro de son auteur. Retrouver tout ce qu'une personne a écrit revient donc à chercher son numéro dans cette table. Dans une base de *documents*, chaque fiche est un petit arbre de champs et de sous-champs. Le format XML (*eXtensible Markup Language*) délimite chaque champ par des *balises*, comme : `<naissance>1320</naissance>`. Le format JSON (*JavaScript Object Notation*) note la même chose en paires *clé-valeur*, un nom de champ suivi de son contenu, `"naissance": 1320`, et une valeur peut elle-même contenir d'autres paires, d'où l'arborescence. Dans les deux cas, deux fiches peuvent ne pas avoir les mêmes champs. Une base de données *graphe* représente l'information comme un réseau : des nœuds pour les personnes, les lieux ou les institutions, des arêtes pour leurs relations. Le format RDF (*Resource Description Framework*), enfin, est également un modèle de données en graphe. Il représente l'information par des triplets sujet-prédicat-objet, par

exemple « Jean / est né en / 1320 ». Des identifiants communs permettent de relier des ressources décrites dans plusieurs bases. Une *ontologie* en fixe alors le vocabulaire, la liste des types d'entités et de relations admis et leur sens.

Chaque famille répond à un compromis différent. Aux tableurs et gestionnaires de fichiers des débuts ont succédé les bases relationnelles [2]. Elles imposent aux données un schéma commun, sur lequel s'appuient les requêtes, leur rapidité et leur précision dépendant ensuite du schéma retenu, des index et de la formulation des requêtes. Les représentations documentaires en XML ou JSON permettent de conserver des fiches dont la structure varie. Cette souplesse répond à la diversité des informations prosopographiques, qu'il peut être difficile de prévoir dans un schéma de tables. Studium autorise ainsi le texte libre dans des champs peu structurés, afin de faciliter la saisie des notices et leur recherche en plein texte [2].

Les projets couvrent des mondes très divers, et chaque choix de modélisation s'y explique par l'usage visé. La China Biographical Database [1], qui décrit plusieurs centaines de milliers d'individus de l'histoire chinoise du VII^e^ au XIX^e^ siècle, est une base relationnelle. Le projet conçoit chaque événement d'une vie comme une relation entre entités. Une nomination relie une personne, un office et une juridiction, ce que le modèle relationnel capture directement. Il destine ses données à l'analyse statistique, spatiale et de réseaux [15], et un système d'analyse visuelle bâti sur cette base croise chronologies, cartes et réseaux de relations pour rendre visible la part d'incertitude de ses événements, dont près de 60 % n'ont ni date ni lieu [79]. La Digital Prosopography of the Roman Republic [2], se concentrant sur la République romaine, publie toutes ses données en RDF, alors même que sa base de travail est relationnelle. L'argument avancé est qu'une prosopographie relie par nature des sources multiples autour des mêmes personnes, et que cette publication ouverte permet à d'autres projets d'interroger et de croiser le résultat comme des données, quand des pages web ne se prêtent qu'à la lecture [10]. Deux bases sont plus proches, par leurs objets, du monde universitaire dont relève Studium. L'Onomasticon de l'université de Pérouse [3] recense les enseignants et les étudiants de cette université depuis le XIV^e^ siècle. Chaque personne y est reliée à ses grades, à ses enseignements et à ses variantes de nom notamment. Ce dernier point pèse, puisque 7 700 personnes sont connues sous plus de 21 000 formes, la même personne apparaissant sous des graphies différentes selon les documents. Son histoire technique passe par deux de ces familles. Lancée en 2008 sur une base relationnelle, l'application a été refondue en 2015 sur un modèle de graphe, les entités devenant des sommets et leurs associations des arêtes, la refonte se réclamant explicitement du modèle factoïde du King's College de Londres [60]. Le projet e-NDP [4] part des vingt-six registres de délibérations du chapitre cathédral de Notre-Dame de Paris (1326–1504), la communauté de chanoines qui administrait la cathédrale et son cloître. Les registres sont d'abord transcrits par reconnaissance automatique de l'écriture manuscrite puis corrigés de manière collaborative, et les personnes mentionnées dans les délibérations alimentent une base nominative. Cette base est relationnelle, servie par une interface de programmation ouverte [5]. Une fiche peut, de

1. https://cbdb.hsites.harvard.edu/ (consulté le 01/09/2026)
2. https://kdl.kcl.ac.uk/projects/dprr/ (consulté le 01/09/2026)
3. https://onomasticon.unipg.it/ (consulté le 01/09/2026)
4. https://endp.chartes.psl.eu/ (consulté le 01/09/2026)
5. https://github.com/chartes/endp-person-app (consulté le 01/09/2026)

plus, pointer vers la page que d'autres bases consacrent à la même personne, Wikidata ou Studium Parisiense par exemple. Ce lien permet au lecteur de consulter les informations que ces dernières consacrent au même chanoine.

Une famille entière de bases partage enfin un même modèle, sous le nom de Factoid Prosopography Ontology[6]. La Prosopography of Anglo-Saxon England[7] recense les habitants de l'Angleterre de la fin du VI^{e} au XI^{e} siècle, la Prosopography of the Byzantine World[8] les individus nommés dans les sources de l'empire byzantin, People of Medieval Scotland[9] les personnes des chartes écossaises médiévales. Le modèle circule aussi hors de ces projets. SPEAR (*Syriac Persons, Events, and Relations*)[10] décrit les personnes que mentionnent les chroniques, les lettres et les vies de saints écrites en syriaque, la langue du christianisme d'Orient. Le projet encode ses factoïdes en TEI (*Text Encoding Initiative*)[11], un format de balisage employé pour décrire la structure et le sens d'un texte, et les publie aussi sous forme de triplets [68].

L'approche par factoïdes, introduite au chapitre 1, organise les informations en assertions sourcées [11]. Elle permet de conserver les affirmations de plusieurs sources, y compris lorsqu'elles se contredisent, et prévoit de leur associer un degré de fiabilité. Cette structuration facilite des recherches croisées, par exemple sur les titres portés par les personnes impliquées dans un type d'événement [56]. DAPHNÉ, le modèle du chapitre 1, reprend cette approche parce qu'il se veut générique et capable de décrire la plupart des bases existantes.

2.2 Les graphes de connaissances

Un *graphe de connaissances* applique à un domaine entier la représentation en réseau décrite plus haut, des arêtes typées pour les relations : « a étudié à », « a enseigné à », « est né à ». Le tout est généralement adossé à une ontologie qui fixe les types admis [30]. Le terme s'est répandu après que Google a donné ce nom, en 2012, au réseau d'entités qui complète son moteur de recherche, et de grands graphes généralistes comme Wikidata ou DBpedia servent depuis de références communes pour désigner les mêmes personnes, lieux ou institutions d'une base à l'autre [30]. Le chapitre 3 utilise Neo4j pour stocker le graphe construit pendant le stage. La publication de données en RDF constitue une autre possibilité d'exploitation, présentée ici comme une perspective. La figure 2.1 illustre le principe d'un graphe de connaissances avec quatre entités.

Une représentation en graphe rend explicites les relations mobilisées par certaines requêtes prosopographiques, « qui a étudié à Paris, sous quels maîtres » se formulant en un parcours d'arêtes. Les mêmes informations peuvent être traitées dans d'autres modèles. Le choix du graphe organise ici leur représentation autour des entités, des factoïdes et de leurs

6. https://www.kcl.ac.uk/factoid-prosopography/projects (consulté le 01/09/2026)
7. https://pase.ac.uk/ (consulté le 01/09/2026)
8. https://pbw2016.kdl.kcl.ac.uk/ (consulté le 01/09/2026)
9. https://www.poms.ac.uk/ (consulté le 01/09/2026)
10. https://spear-prosop.org/ (consulté le 01/09/2026)
11. https://tei-c.org/ (consulté le 01/09/2026)

liens, chaque assertion sourcée du modèle factoïde devenant un morceau de graphe qui relie une personne, un fait et une source. Le web sémantique prolonge cette démarche en publiant des données structurées et en les reliant d'un site à l'autre. Hyvönen distingue trois usages qu'en font les humanités numériques, la publication de corpus en triplets que l'on cherche et parcourt, leur exploitation comme matériau de recherche que l'on interroge et croise d'une collection à l'autre, et des portails capables de repérer eux-mêmes des questions, d'y répondre et d'expliquer leur raisonnement, le troisième n'étant encore, dans son article, qu'une perspective [34]. La conversion de Studium relève du premier et prépare le deuxième, transformer des fiches faites pour être lues en données que l'on interroge et c'est sur ces deux-là que les modèles de langage ont d'abord porté. Des systèmes récents combinent traitements déterministes et modèle de langage, les règles traitant le gros des données et le modèle servant de repli pour les cas sémantiquement difficiles, ce qui contient les coûts sans sacrifier l'exactitude [43]. Le pipeline du chapitre 3 est de cette famille, des règles pour presque tout, un modèle de langage pour le seul typage des institutions. Construire un graphe ainsi ajoute cependant une incertitude à celle des sources, qu'il faut repérer et traiter [35].

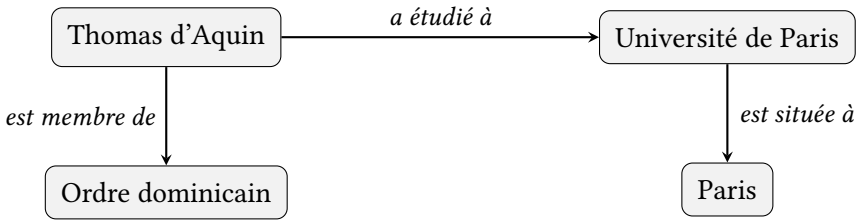

Figure 2.1 – Un graphe de connaissances réduit à quatre entités et trois relations. La question « quels dominicains ont étudié dans une université parisienne ? » peut être traduite en un parcours reliant l'ordre religieux, les personnes, l'université et sa ville.

2.3 Les formes du doute

Le terme « incertitude » recouvre plusieurs phénomènes selon les auteurs et les disciplines : manque de connaissance, doute, ambiguïté, flou, subjectivité, erreur ou risque [55]. Pour l'analyse des notices, nous considérons qu'une phrase exprime une incertitude au sens large lorsqu'elle signale explicitement un doute, une approximation ou un manque d'information sur un élément. Les chapitres 4 et 5 précisent ensuite les conventions retenues pour leurs tâches respectives. Les phénomènes ainsi couverts restent divers, et la littérature s'attache à les distinguer. La figure 2.2 résume en quatre questions les distinctions que cette section détaille. Pourquoi l'énoncé n'est-il pas sûr, et comment le texte le dit-il ? Ces deux premières questions correspondent aux deux plans que sépare la typologie de Vincze, construite pour la détection automatique en anglais et en hongrois [75]. Au plan *sémantique*, la question est de savoir pourquoi la proposition ne peut pas être tenue pour simplement vraie, et quatre réponses définissent quatre types. Épistémique, quand on ignore si le fait est vrai au moment où il est énoncé, « il peut pleuvoir ». Doxastique, quand une croyance est rapportée, « il croit que la Terre est plate ». D'investigation, quand le fait est en cours d'examen, « nous étudions le rôle de cette protéine ». Conditionnel, quand la vérité dépend

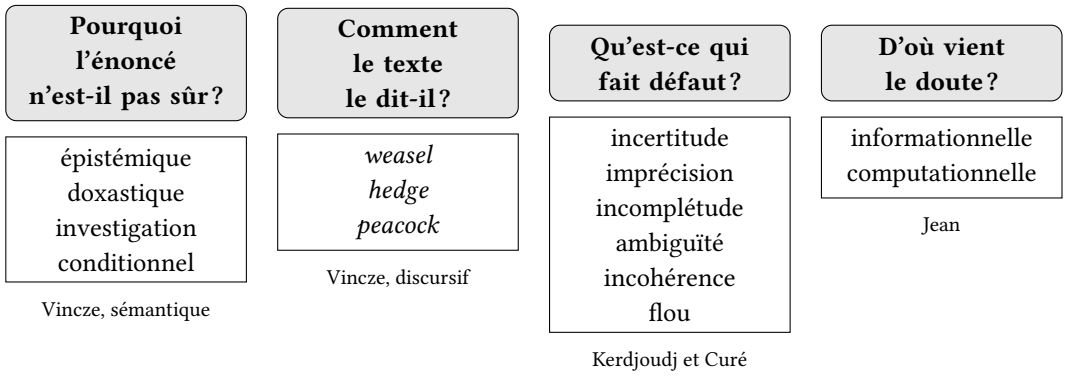

Figure 2.2 – Les typologies de l'incertitude ramenées à quatre questions posées à un même énoncé. Sur « il obtint peut-être sa maîtrise vers 1350 », le doute est épistémique, dit par des *hedges*, il mêle un fait incertain et une valeur imprécise, et son origine est informationnelle.

d'une condition, « s'il pleut, nous restons à l'intérieur ». Au plan *discursif*, la question n'est plus ce que le texte affirme mais la manière dont il le formule, en trois catégories héritées des conventions éditoriales de Wikipédia. Les expressions de type *weasel* laissent la source indéterminée : « certains disent que », « il paraît que ». Les *hedges* sont des formulations atténuées ou approximatives, « peut-être », « environ », « probablement ». Les expressions de type *peacock* évaluent subjectivement une information, « remarquable », « controversé ». Cette grille aide à décrire les notices. Une formule comme « peut-être » peut signaler un doute épistémique sur le fait rapporté. « Selon X » ou « d'après X » attribue une affirmation à un tiers ; cette attribution ne suffit pas, à elle seule, à établir que le rédacteur en doute. Au plan discursif, les notices comportent notamment des *hedges*, comme « vers » ou « peut-être », et des expressions de type *weasel*, comme « on dit que » ou « il passe pour ».

Le *hedge*, dominant dans les notices, vient de l'étude du *hedging* scientifique, c'est-à-dire la manière d'atténuer une affirmation pour ne pas s'engager au-delà de ce que les données autorisent, « ces résultats *suggèrent* que… ». Hyland en recense les procédés [33], et la typologie usuelle remonte à Prince, Frader et Bosk [59]. Les *shields*, littéralement les boucliers, éloignent l'affirmation de l'auteur, par la plausibilité, « probablement », « il semble que », ou par le report sur un tiers, « selon Duchesne », sans dire si l'auteur y adhère. Les *approximateurs* rendent floue la valeur plutôt que l'énoncé, « environ », « autour de ». Salager-Meyer adapte cette typologie au discours médical écrit [66]. La liste de marqueurs du chapitre 5 en descend directement, et le dictionnaire repris au chapitre 4 procède du même fonds.

Qu'est-ce qui fait défaut à l'information ? Plusieurs cas se laissent distinguer. L'information est *incertaine* quand le fait lui-même est douteux, « il a peut-être enseigné à Paris ». Elle est *imprécise* quand le fait est acquis mais sa valeur approximative, « il fut reçu maître vers 1350 ». Elle est *incomplète* quand elle manque, un champ réduit à « ? ». Kerdjoudj et Curé recensent ces trois défauts et trois autres parmi les imperfections des textes du web, avant de ne traiter que l'incertitude [38]. L'information est *ambiguë* quand elle admet plusieurs lectures que le texte ne départage pas, « Pierre a dit à Paul qu'il avait gagné ». Elle est *floue* quand elle est imprécise sans que ses bornes soient données, « avant 1347 » fixe une borne, « vers 1350 » n'en fixe aucune. Elle est enfin *incohérente* quand deux énoncés

ne peuvent être vrais ensemble, « il est né en 1350 », et « il s'est marié en 1348 ».

D'où vient le doute, enfin ? L'incertitude est dite *informationnelle* quand la source doute elle-même, *computationnelle* quand c'est le traitement automatique qui a pu se tromper [36].

Deux cadres plus généraux combinent ces axes. L'ontologie URREF (*Uncertainty Representation and Reasoning Evaluation Framework*) vient de la fusion d'informations, le domaine où un système combine les rapports de plusieurs sources, des capteurs ou des informateurs, pour établir une situation [18]. Elle standardise l'évaluation de ces systèmes en posant quatre questions à toute information incertaine. Le doute vient-il du hasard du monde ou d'un manque de connaissance ? Qu'est-ce qui fait défaut, ambiguïté, incomplétude, imprécision, aléa ou incohérence ? Le degré a-t-il été mesuré ou jugé ? Et quel formalisme le traitera, probabilités, ensembles flous ou fonctions de croyance ? Ce cadre sert ici à situer les dimensions de l'incertitude. Le stage porte plus directement sur l'étape qui consiste à produire une annotation à partir du texte. L'inventaire des formulations de l'incertitude scientifique offre, à cet égard, un point de comparaison : il distingue douze groupes, dont les expressions conditionnelles, les hypothèses, les prédictions et les marques de subjectivité, décrits par des patrons linguistiques [55].

Cet inventaire a toutefois été construit pour des textes scientifiques contemporains en anglais. Les notices étudiées sont rédigées en français et utilisent des conventions de saisie. Nous avons donc examiné le corpus pour définir les annotations utilisées au chapitre 5.

Un autre schéma, repris dans la figure 2.3, situe l'incertitude des traitements automatiques sur trois axes [62, p. 79–80]. Chaque axe éclaire un aspect de Studium. L'*exactitude de l'information* d'abord. Le doute qu'un rédacteur glisse en texte libre peut relever de chacune de ses cases, une source jugée peu fiable, un fait dont la véracité est discutée, deux sources qui se contredisent, un commentaire libre qui nuance. Les deux *moyens de diffusion* de l'incertitude décrits sont présents dans Studium, avec d'une part des expressions d'incertitude « probablement », « peut-être », et d'autre part des métadonnées, comme un « ? » écrit au début pour signaler le doute sur tout le reste de la phrase. Le *champ d'usage* de l'incertitude, enfin, situe nos données dans le cadre d'un champ spécifique, sur de la prosopographie médiévale.

2.4 La place de l'incertitude dans les bases prosopographiques

Avant de remplir les attributs de certitude de DAPHNÉ, encore faut-il savoir comment les bases existantes accueillent le doute. L'imprécision des dates, omniprésente dans les sources historiques, en donne une première mesure. Matoušek et al. en recensent sept types d'assertions temporelles, allant de la date précise à la date manquante, en passant par la date incertaine donnée entre deux bornes ou aux alentours d'un jour [51]. Au-delà des dates, Akoka et al., dont le modèle conceptuel a été éprouvé sur trois bases dont Studium, indiquent en 2021 ne pas avoir identifié de base prosopographique représentant l'incertitude au niveau du modèle [3]. Ce constat motive leur proposition de la représenter explicitement. Les réserves peuvent être conservées dans le texte libre, dans des notes ou dans des

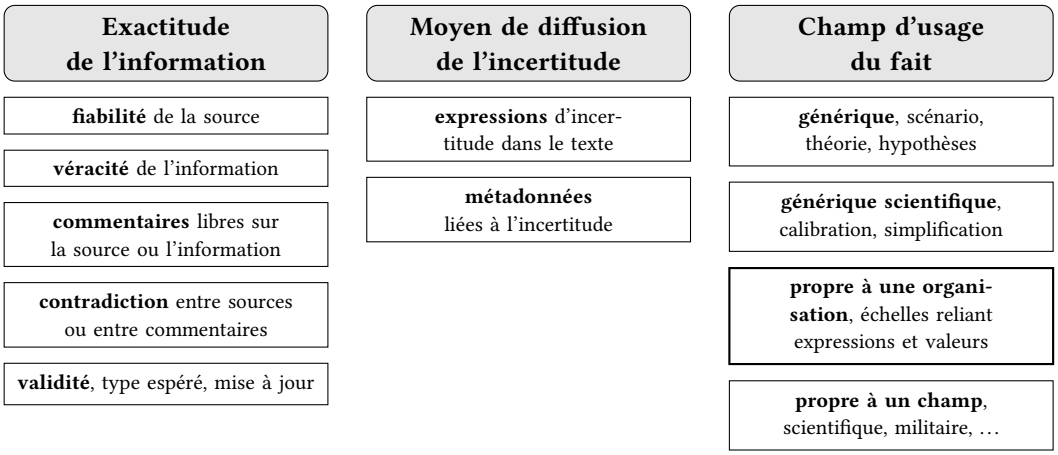

Figure 2.3 – Le schéma des dimensions de l'incertitude des traitements automatiques, repris de [62, p. 79–80].

attributs ajoutés aux données [49]. Le standard TEI prévoit également des mécanismes pour annoter la certitude et la précision dans un texte [72]. Le modèle DAPHNÉ, lui, distingue la certitude de l'information, la confiance dans la source et la précision de la valeur. Une fois ces attributs renseignés, des interrogations précises peuvent ainsi avoir lieu, avec des réponses qui peuvent être appuyées par des sources dont la fiabilité dépasse un seuil [3].

Dans un second article, Akoka et al. décrivent le chemin qui mènerait des sources aux faits établis dans une telle base [4]. L'article distingue quatre objets. La *mention* est une information issue d'un fragment de source. Sa crédibilité vaut 1 quand l'information est affirmée sans réserve et prend sinon une valeur strictement comprise entre 0 et 1. Le *factoïde*, dans un sens propre à cet article, s'appuie sur une ou plusieurs mentions qui rapportent la même information. Le *fait* est un factoïde tenu pour acquis par la communauté historiographique, au besoin après réconciliation de plusieurs factoïdes voisins dont il ne retient que la part commune. Enfin, l'*hypothèse* est une conjecture que l'historien formule sur la relation entre plusieurs faits ou factoïdes. Le processus proposé enchaîne cinq phases sur ces objets (figure 2.4). La capture extrait les mentions des sources avec leur crédibilité. L'association les regroupe en factoïdes. L'établissement des faits arbitre entre factoïdes au moyen de règles qui agrègent leurs crédibilités. Le test confronte les hypothèses de l'historien aux faits établis. La dernière phase répercute enfin le verdict en amont, la fiabilité d'une source augmentant quand les hypothèses qu'elle appuyait se confirment et diminuant dans le cas inverse. Une partie de notre travail se situe lors de la première phase, celle sur laquelle repose l'entièreté du processus.

2.5 Échelles et niveaux de certitude

Aucune granularité ne fait consensus pour graduer la certitude. La littérature a retenu deux niveaux quand la tâche se réduit à séparer le certain de l'incertain, comme dans le corpus biomédical BioScope et la campagne d'évaluation CoNLL-2010 [76, 21]. D'autres travaux graduent en trois [32], quatre [64], cinq niveaux [63], voire davantage [50], ou sur

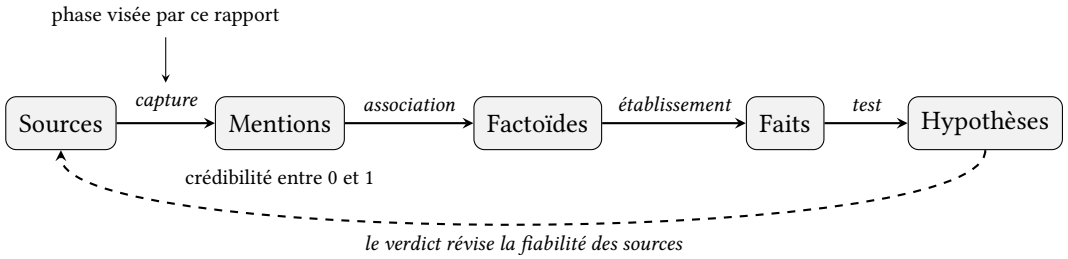

Figure 2.4 – Le processus d'Akoka et al. [4]. Quatre phases mènent des sources aux hypothèses, la cinquième revient en pointillé réviser la fiabilité des sources d'après le verdict du test.

une échelle continue [65]. Ce dernier choix répond à des jugements humains trop divergents pour être forcés dans des catégories fixes [47]. Ces travaux ne graduent d'ailleurs pas tous le même objet, tantôt la spéculation qu'un marqueur signale, tantôt l'engagement de l'auteur, tantôt la factualité de l'événement rapporté, mais le choix d'une granularité s'y pose dans les mêmes termes. Les chapitres 4 et 5 retiennent chacun une échelle à cinq niveaux, de graduations voisines mais distinctes, la première reprise de l'article que le chapitre prolonge, la seconde arrêtée sur le corpus.

La manière d'attribuer un niveau reste le plus souvent implicite, par analogie avec des exemples annotés. Le corpus FactBank fait exception [67]. Chaque événement y reçoit une modalité et une polarité, la modalité vaut certain, probable ou possible, avec une valeur sous-spécifiée quand la source ne se prononce pas, la polarité dit si l'événement a eu lieu ou non. Transposé aux notices, « il fut reçu maître ès arts en 1380 » serait certain et positif, « il n'a probablement pas enseigné à Paris » probable et négatif. La valeur n'est pas absolue, elle est relative à une source, l'auteur du texte par défaut, ce qui permet de noter qu'un texte rapporte les certitudes d'un tiers sans les endosser. L'ordre possible $<$ probable $<$ certain s'appuie sur les tests d'échelle de Horn [31]. Le principe est qu'un terme faible peut s'affirmer en laissant la porte ouverte au plus fort, jamais l'inverse. « C'est possible, si ce n'est probable » se dit, « c'est probable, si ce n'est possible » ne se dit pas. De là découle un test que l'annotateur peut réellement appliquer [67], qui consiste à enchaîner l'énoncé avec son contraire et à juger si le résultat heurte. « Il se peut que cela ne change pas, mais cela changera probablement » passe, donc le premier énoncé relevait du possible. « Cela ne changera probablement pas, mais cela changera probablement » se contredit, donc probable. Ce procédé rend le jugement de l'annotateur plus explicite, tout en lui laissant l'appréciation de l'enchaînement. Mais reformuler chaque énoncé puis juger le résultat a un coût difficile à tenir sur des milliers de phrases, et nous ne l'avons donc pas retenu.

Une autre manière de rendre l'attribution explicite fixe les valeurs dans le lexique lui-même. L'analyse du renseignement en fournit un exemple avec la table de Kent, du nom de l'analyste qui voulait harmoniser les estimations du renseignement américain. Celle-ci fait correspondre à chaque expression une probabilité assortie d'une marge (tableau 2.1) [37]. Kent donne pour « probable » un intervalle, 75 ± 12, soit 63 à 87 %. Cette représentation trace une frontière nette entre les valeurs incluses et exclues, sans préciser si certaines correspondent mieux que d'autres au terme « probable ». Les expériences de Wallsten et ses collègues suggèrent une lecture plus graduelle de ces expressions, à l'aide de *fonctions*

d'appartenance floues [77]. Une telle fonction associe à chaque probabilité un degré d'appartenance compris entre 0 et 1, qui indique à quel point cette valeur correspond à l'expression étudiée. Ce degré d'appartenance et la probabilité sont donc deux nombres de sens différent. L'étude met également en évidence des différences entre les interprétations individuelles, une même probabilité ne correspond pas au même degré à « probable » pour tous les participants. Les seuils proposés par Kent ne reflètent donc pas nécessairement la manière dont chaque lecteur comprend ce terme.

Expression	Probabilité
certain	100 %
presque certain	93 ± 6 %
probable	75 ± 12 %
chances à peu près égales	50 ± 10 %
probablement pas	30 ± 10 %
presque certainement pas	7 ± 5 %
impossible	0 %

Table 2.1 – La table de Kent, traduite de [37]. Chaque expression est associée à une probabilité centrale et, le cas échéant, à un intervalle proposé par Kent.

Kerdjoudj et Curé, effectue un mouvement similaire pour l'étude de textes du web, car chaque marqueur y reçoit un degré élevé, modéré ou faible, 0,75, 0,50 ou 0,25, qu'un modificateur comme « très » ou « moins » décale de 0,15. Les incertitudes qui pèsent sur une même information extraite, un marqueur, un discours rapporté, un verbe au futur, se combinent enfin par réseau bayésien en un degré de confiance unique [38].

La difficulté précède d'ailleurs les mots, car sur une situation singulière le jugement de probabilité est lui-même difficile à former et, a posteriori, à vérifier. Le bureau que dirigeait l'auteur de la table jugeait improbable, en septembre 1962, l'installation de missiles soviétiques à Cuba, découverts le mois suivant, quand le président américain estimait au plus fort de la crise entre une chance sur trois et une sur deux que l'Union soviétique aille jusqu'à la guerre [71]. Cette question, traitée depuis longtemps [23] déborde le cadre de ce rapport.

2.6 Marqueurs et portée

La détection automatique peut porter sur plusieurs unités : déterminer si une phrase exprime une incertitude, repérer les mots qui la signalent, puis délimiter leur portée. Les objectifs et les méthodes sont similaires pour la négation, et les deux tâches sont souvent traitées ensemble [46]. Le corpus BioScope de Vincze et al. annote les marqueurs d'incertitude et de négation ainsi que leur portée dans des textes biomédicaux anglais [76]. La campagne CoNLL-2010, organisée par Farkas et al., s'appuie notamment sur ce corpus pour évaluer la détection de l'incertitude et de sa portée [21]. Le français dispose d'équivalents biomédicaux, les corpus CAS et ESSAI [19].

Trois générations de méthodes s'y sont succédé [55]. Les premières reposent sur des

règles, un lexique de marqueurs assorti de déclencheurs contextuels qui écartent les emplois non incertains [13], ou des motifs lexico-syntaxiques pour délimiter la portée [5]. L'apprentissage supervisé a suivi, de l'amorçage faiblement supervisé de Medlock et Briscoe [52] aux systèmes de CoNLL-2010, qui classent chaque mot d'après ses traits de surface, sa catégorie grammaticale et ses dépendances syntaxiques [21]. Les transformers ont pris le relais et y ont établi les meilleurs résultats [39]. Ces détecteurs sont néanmoins spécialisés par langue et par domaine et se transfèrent mal [75].

Dans tous ces travaux, la portée reste un morceau de texte. Sur « il a peut-être enseigné à Paris », ils délimiteraient les mots que le « peut-être » couvre, soit le segment « enseigné à Paris ». Pour la tâche étudiée ici, il faut également relier cette portée textuelle aux objets du graphe. Personne ne doute de l'existence du personnage ni de celle de Paris, c'est le lien entre les deux que le « peut-être » affaiblit. La réponse attendue n'est donc plus un segment mais un objet, l'arête ou le factoïde qui enregistre l'enseignement, et aucun de ces corpus n'annote ce niveau-là.

Un même marqueur n'exprime en outre pas toujours un doute, « pouvoir » disant tantôt l'éventualité et tantôt la capacité ou la permission, et cette polysémie borne les approches par simple lexique [55]. L'étude de Kouadri et al. [40] illustre cette difficulté dans les notices historiques et restreint le patron retenu à « il se peut que ». Pour expliciter la distinction, on peut comparer deux exemples construits : « il se peut qu'il ait assisté au procès » exprime une possibilité, tandis que « il peut lire le grec » décrit une capacité.

2.7 L'incertitude dans les textes scientifiques et historiques

L'incertitude exprimée dans les textes a été étudiée sur des terrains variés. D'une part, sur les pages du web, où l'extraction de connaissances rencontre des forums, des blogs ou de la presse qui mêlent l'affirmé et le douteux [38]. D'autre part, dans les textes scientifiques, où un chercheur dose ses affirmations parce que ses résultats, une expérience, une mesure, une statistique, restent soumis à interprétation, et ce langage a ses corpus d'étude et ses détecteurs dédiés [6, 54, 28]. Les notices partagent le geste de ces textes, un rédacteur chercheur qui atténue par des marqueurs, mais pas l'objet, leur doute porte sur des faits que des sources lacunaires attestent mal, non sur le sens à donner à une expérience. Leur terrain propre est celui de l'histoire et de l'archéologie, où représenter le doute des données est reconnu comme un problème de modélisation en soi, du temps historique incertain [42] aux cadres de logique floue proposés pour le qualifier et le quantifier [50]. Ces propositions restent peu reprises au-delà de leurs projets d'origine, les approches statistiques traitant le flou comme une marge d'erreur à réduire, quand il est ici une propriété de l'information elle-même [49]. Ces derniers auteurs distinguent deux origines au flou de ces données. Le *vagueness* épistémique tient à l'ignorance du chercheur, comme la localisation d'une cité ancienne ou l'appartenance d'un document à une période, et recule quand la documentation s'enrichit. Le *vagueness* ontologique tient aux objets eux-mêmes, dont les frontières restent floues quelle que soit la finesse de la mesure, celles d'une montagne par exemple [49]. Le doute des fiches de Studium relève du premier, un registre retrouvé pourrait le lever. Le

second loge plutôt dans les catégories que la base fige par convention, le contour d'une discipline ou d'un statut, et reste hors du champ de ce travail.

Sur Studium même, Kouadri et al. ont publié une première étude de détection d'incertitude déjà citée [40]. Elle repose sur un échantillon de 896 fiches, soit 53 537 champs dont un peu plus de la moitié en texte libre. L'étude classe le doute des notices en trois espèces. La première tient à un mot du lexique, la deuxième à une construction reconnaissable, la troisième au seul contexte de la phrase, « il y a débat sur sa présence à l'université de Paris » n'annonçant son incertitude par aucun marqueur isolable. La méthode s'attaque aux deux premières. Un dictionnaire français de l'incertitude de 190 entrées, dont le chapitre 4 détaille la construction à partir de WoNeF [58], couvre les mots du lexique. Des patrons dégagés à la lecture des notices couvrent les constructions, dates et lieux approchés (« aux alentours de 1550 »), alternatives entre hypothèses, valeurs signalées comme estimées, conditionnel, forme interrogative, et la tournure « il se peut que » rencontrée plus haut. L'évaluation chiffre l'apport de chacun. Le dictionnaire seul obtient 95 % de précision, mais ne retrouve que 24 % des cas incertains. Les patrons seuls atteignent 48 % de rappel. Leur combinaison porte le rappel à 59 % pour 89 % de précision, soit un F1 de 71 %. Sur ce jeu, la combinaison repère donc davantage de cas, au prix d'une augmentation de faux positifs.

La troisième espèce, celle du seul contexte, échappe par construction aux dictionnaires comme aux patrons, et les auteurs concluent qu'il y faudrait de l'apprentissage automatique, tout en relevant que celui-ci réclame de grandes quantités de données annotées et que cette annotation coûte cher. L'apprentissage en contexte, présenté à la section suivante, répond autrement à cette contrainte, un modèle généraliste recevant la règle en consigne, dans le *prompt*, au lieu de l'induire d'un corpus annoté.

2.8 Modèles de langage et apprentissage en contexte

Les modèles de langage sont des réseaux de neurones entraînés sur de vastes corpus, soit à retrouver des mots masqués dans un texte, comme l'encodeur CamemBERT pour le français [48], soit à prédire la suite d'un texte, comme les modèles génératifs interrogés dans ce rapport. Ils peuvent être adaptés à une tâche selon deux stratégies.

La première, le *fine-tuning*, consiste à poursuivre l'entraînement sur un corpus annoté de la tâche, ce qui requiert de disposer d'un jeu d'entraînement assez grand pour que le modèle ne se contente pas de mémoriser les exemples, mais qui permet d'apprendre les conventions d'annotation par la pratique. La seconde, l'*apprentissage en contexte*, consiste pour le modèle à exécuter une tâche à partir d'une consigne en langue naturelle (le *prompt*) et d'une poignée d'exemples (le *few-shot*), sans aucun entraînement supplémentaire [12]. Pour noter la certitude d'une phrase de notice, la consigne décrit l'échelle, donne quelques phrases déjà notées, et le modèle note les suivantes. Le paradigme est attractif pour la détection d'incertitude, où les corpus annotés sont rares et coûteux à construire, en français historique plus qu'ailleurs. Il gagne aussi la construction de graphes de connaissances elle-même. Li y décrit le passage du modèle entraîné pour une seule tâche au modèle générique piloté par consigne, et soutient que, sur l'extraction de relations, un modèle instruit sur la tâche peut rivaliser avec les modèles entraînés sur données annotées, à condition que la

consigne encode soigneusement la tâche et le schéma attendus [43]. L'effort se déplace aussi, la performance dépendant de la consigne autant que du modèle, du format et du choix des exemples [80, 45]. La comparaison du chapitre 4 oppose ainsi un encodeur, CamemBERT, adapté par apprentissage supervisé sur les phrases annotées, à des modèles généralistes qui appliquent un barème explicité dans la consigne.

Une partie de la littérature des modèles de langage examine si les modèles savent estimer leur *propre* incertitude, en la verbalisant ou en donnant sur demande un score de confiance, et si ces degrés correspondent à leurs taux de réussite réels, ce que mesure la *calibration*. Le constat récurrent est une surconfiance des scores bruts, que des stratégies de sollicitation corrigent en partie [73], un modèle pouvant par ailleurs apprendre à verbaliser une incertitude calibrée [44]. La question complémentaire, savoir si les modèles détectent et quantifient fidèlement l'incertitude qu'un auteur a délibérément inscrite dans son texte, est moins abordée et n'est presque pas explorée pour le français traitant de données historiques.

2.9 Les liaisons incertaines

Décider si deux notices décrivent la même personne est un problème formalisé de longue date sous le nom de *couplage d'enregistrements*. Il est automatisé dès 1959, chaque attribut étant pesé par sa rareté, un nom rare qui concorde valant plus qu'un nom fréquent [53], et Fellegi et Sunter en donnent la théorie [22]. Chaque paire de fiches y est comparée attribut par attribut, et chaque concordance ou discordance pèse selon sa fréquence chez les paires qui désignent réellement la même personne et chez les autres. Leur réponse à la question « combien d'attributs partagés suffisent » est un rapport de vraisemblance, avec deux seuils dérivés des taux d'erreur tolérés et trois issues : le lien, le non-lien, et une zone d'indécision qui renvoie formellement le cas à l'expertise humaine.

Tversky en fournit la lecture ensembliste [74]. Chaque fiche devient un ensemble de traits, un nom, un lieu, une date. Soient A et B les ensembles de traits de deux fiches, $A \cap B$ leurs traits communs et $A \setminus B$ les traits que la première est seule à porter. La similarité s'écrit

$$s(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha\,|A \setminus B| + \beta\,|B \setminus A|},$$

où $|\cdot|$ désigne le cardinal et où les coefficients α et β règlent le poids accordé aux traits propres. Prendre $\alpha = \beta = 1$ redonne l'indice de Jaccard, la part des traits partagés parmi tous les traits observés, $J(A,B) = |A \cap B| \,/\, |A \cup B|$, et $\alpha = \beta = \frac{1}{2}$ celui de Dice, $D(A,B) = 2\,|A \cap B| \,/\, (|A| + |B|)$. La théorie des ensembles approximatifs, les *rough sets*, délimite ce qu'une telle comparaison peut décider. L'identité ne s'y juge que relativement aux attributs disponibles, deux fiches qui coïncident sur chacun d'eux sont dites indiscernables, et l'indiscernabilité n'est pas l'identité [57]. La ressemblance n'est enfin pas transitive, car les erreurs s'additionnent [70].

Les historiens pratiquent ce couplage depuis les registres paroissiaux [78], et il est possible de reconstruire ainsi des populations entières à partir de données rares, ambiguës et fautives [9]. Un dilemme du seuil se pose alors. Exiger beaucoup avant de lier évite les

fausses identités mais laisse la même personne éclatée en plusieurs fiches, exiger peu lie davantage mais lie à tort. Les travaux récents mesurent les deux, la précision des liens posés et la part des vrais liens retrouvés, et règlent leurs seuils en connaissance de cause [17, 1]. Ces cadres font aussi leur place aux erreurs de saisie. Une discordance ne prouve pas que deux fiches désignent deux personnes, car la graphie a pu dévier ou la date être mal copiée, et le modèle chiffre cette possibilité, soit la probabilité qu'une paire désignant bien la même personne diverge malgré tout sur un attribut. Mais l'incertitude y naît de la comparaison entre deux fiches, aucune valeur n'arrivant porteuse de son propre degré de certitude. Des extensions existent de part et d'autre. Le couplage a été étendu aux bases de données probabilistes, où les valeurs d'attributs ou les enregistrements eux-mêmes portent une probabilité [7], et il a été appliqué à des textes médiévaux et modernes, chartes et archives où les personnes ne sont pas systématiquement identifiées [27]. À notre connaissance, aucun travail ne combine les deux, un couplage sur corpus médiéval où chaque valeur arriverait déjà porteuse de son degré de certitude. Ces questions ne sont pas évaluées dans le stage, mais en constituent une perspective de recherche toute naturelle.

2.10 Bilan du chapitre

De cet état de l'art, nous retenons des distinctions pour l'étude de Studium. L'une sépare le statut de l'information de sa formulation : un même doute peut être exprimé par « peut-être », par une attribution à une source ou par une convention telle que « ? ». Une autre distingue l'incertitude, qui porte sur la véracité d'un fait, de l'imprécision, qui affecte sa valeur, par exemple dans une date introduite par « vers ». La dernière oppose l'incertitude présente dans les sources à celle que produit le traitement automatique. Ces distinctions orientent l'analyse, sans fournir directement un schéma d'annotation, car celui-ci doit aussi tenir compte du français des notices, de leur caractère historique et des conventions propres à Studium.

Ce schéma doit aussi trouver sa place dans le modèle. Le factoïde de DAPHNÉ prévoit un attribut de certitude sur chaque relation, que les bases prosopographiques ne renseignaient pas encore. Le remplir suppose de relier ce que le texte exprime à ce que le graphe représente. Or les corpus de détection annotent un marqueur et une portée qui restent un segment de texte, alors que le graphe attend un objet, l'arête ou le factoïde à affaiblir, et aucun d'eux n'annote ce niveau. Le degré à y inscrire n'est pas mieux fixé, aucune granularité d'échelle ne faisant consensus. Sur Studium même, la seule étude publiée repère l'incertitude sans la situer ni la graduer, et bute sur le doute que le contexte seul exprime, faute de données annotées. L'apprentissage en contexte lève cette contrainte en remplaçant le corpus annoté par une consigne. L'élément visé et le degré, eux, sont à fixer sur le corpus, ce que font les chapitres suivants. Le rapprochement d'entités et le raisonnement sur un graphe incertain restent, enfin, des perspectives.

Deuxième partie

Contributions

Chapitre 3

Des fiches au graphe DAPHNÉ

> `il est peut-être $ANDREAS de Suetia$, cité dans le &Computus& de %1329-1330%.`
>
> Studium Parisiense, fiche n° 640, Andreas Freouati

3.1 Ce que les fiches donnent à voir

Une fiche Studium mêle deux sortes de contenu. Le volet `meta` présente les noms, les lieux, les institutions et les dates sous une forme régulière, que des règles explicites convertissent directement en nœuds et en arêtes. Le volet `value` et les commentaires portent au contraire la phrase de l'historien, avec tout ce que la grille ne peut pas accueillir.

Cette première partie du chapitre traite la part structurée, celle qui se convertit par règles. Une partie lui échappe malgré tout. Le champ des institutions mêle des entités de natures trop diverses pour une règle, pour lesquelles nous évaluons une méthode fondée sur un modèle de langage à la section 3.5. Le texte libre, où se logent la nuance et le doute, n'est pas traité ici. Les chapitres 4 et 5 lui sont consacrés.

3.2 Du JSONL au graphe

Le pipeline lit le JSONL et produit un graphe Neo4j qui suit le schéma DAPHNÉ. Il procède en plusieurs temps. Les fiches qui portent le même nom principal, ainsi que les mêmes variantes de noms, sont d'abord ramenées à une seule, la plus récemment mise à jour. Chaque fiche restante est ensuite parcourue champ par champ, et chaque valeur reconnue donne un nœud ou une arête, après nettoyage du texte et lecture des dates. Le résultat est enfin écrit en fichiers CSV, un par type de nœud et d'arête, puis chargé dans Neo4j. Le chargement est déterministe : relancer le pipeline sur les mêmes données redonne le même graphe sans créer de doublons.

Concrètement, le pipeline tient dans un script d'orchestration appuyé sur quatre modules, la configuration, l'extraction, l'écriture des CSV et le chargement. Il prend en entrée l'export JSONL et produit un dossier de fichiers CSV, que le chargeur verse ensuite dans Neo4j. Les deux moitiés sont indépendantes, on peut régénérer les CSV sans toucher à la base, ou recharger la base sans tout réextraire. La figure 3.1 résume ce que chaque étape reçoit et produit.

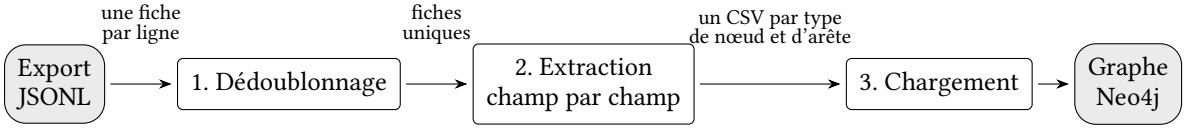

Figure 3.1 – Le pipeline, de l'export JSONL au graphe Neo4j. Les flèches portent ce qui circule entre les étapes, et les CSV intermédiaires permettent de recharger le graphe sans refaire l'extraction.

Les nœuds produits couvrent les principaux types du schéma, PhysicalPerson, Name, GroupP, Place, Zone, Source, Factoid et FactoidType, Time, Domain, Object, ObjectType, Role et Rank, reliés par une trentaine de types d'arêtes. Deux choix de représentation changent la lecture du graphe. Un nœud Time porte en attribut la distinction entre instant et intervalle que le modèle confie à des spécialisations. Le rôle et le rang d'une personne dans un factoïde sont portés en attributs de l'arête Participate, si bien que les nœuds Role et Rank existent sans qu'aucune arête ne les atteigne. Ces choix constituent deux écarts explicites au modèle conceptuel. Chaque factoïde, dont la section 1.2 a exposé le rôle, est rattaché à l'un des treize types retenus, un type par rubrique de la fiche, ce qui conserve pour chaque factoïde la rubrique dont il provient.

Un exemple

La figure 1.4 a montré ce que le modèle attend pour la ligne `datesOfActivity` d'Andreas Freouati. Voici le bloc dont elle est tirée, tel qu'il figure dans l'export.

```
{ "value": "%1329-1348%",
  "meta": { "dates": [ { "type": "INTERVAL",
      "startDate": { "type": "SIMPLE", "date": 1329 },
      "endDate":   { "type": "SIMPLE", "date": 1348 } } ] },
  "comment":   [ "il est peut-être $ANDREAS de Suetia$, cité dans
                 le &Computus& de %1329-1330%." ],
  "reference": [ "CUP: II, n° 1184, p. 671 ; réédition dans
                 COURTENAY, &Parisian Scholars...&, p. 221." ] }
```

Le pipeline ne lit pas le texte « %1329-1348% ». Il lit le volet `meta`, où le projet a déjà déposé un intervalle typé. Le modèle attend de ce bloc un nœud Factoid et cinq arêtes, dont deux ReferTo vers les documents cités dans le volet `reference`. Le pipeline actuel en produit quatre, la seule arête ReferTo pointant vers un nœud Source qui représente la notice entière. Il ne lit pas encore le volet `reference` des blocs, si bien que tous les factoïdes

d'une notice partagent la même source. Le graphe actuel ne permet donc pas de distinguer les affirmations selon les documents cités dans une même notice.

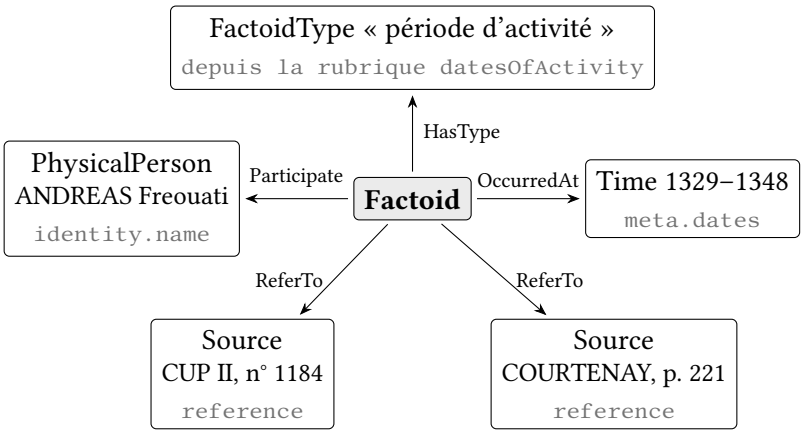

Figure 3.2 – Le même factoïde qu'à la figure 1.4, avec en gris la provenance de chaque branche dans le bloc JSON.

Deux incertitudes que la fiche notait déjà survivent à cette conversion. La première tient à la précision de la date. L'intervalle d'Andreas est net, donc l'arête temporelle porte `datingtype = exact`. Sur une fiche qui écrirait « vers 1329 », la même arête porterait `approximate`, et l'approximation resterait lisible dans le graphe au lieu de devenir une date comme une autre. La seconde est le « ? » que le rédacteur accole à une valeur structurée. Là où `meta.institutions` porte « Uppsala ? », le point d'interrogation est retiré du texte et l'arête correspondante reçoit `certainty = "uncertain"`. Le cas est rare, 26 blocs sur l'export complet. Le « ? » qui tient lieu de valeur entière, plus de 11 000 champs, ne laisse en revanche aucune marque, le pipeline le traitant comme une valeur vide. Les attributs `certainty` prévus par le modèle, dont le chapitre 1 laissait le contenu ouvert, sont ainsi renseignés pour les valeurs portant ce marqueur.

Le pipeline laisse de côté le volet `comment`, et il ne lit le volet `value` que comme une étiquette, pour les grades, les fonctions et la nature des liens familiaux, sans l'analyser. L'attribut `original_text` que le modèle prévoit pour la phrase du rédacteur reste vide, la description du factoïde étant une phrase générée par le pipeline, « Naissance de X ». Sur cet exemple le commentaire porte pourtant l'information la plus délicate de tout le bloc, puisque l'historien y écrit qu'Andreas est *peut-être* le même homme qu'Andreas de Suetia. C'est une hypothèse d'identité, formulée en français, que les règles implémentées ne prennent pas en charge. Rattacher au graphe les annotations des chapitres suivants demandera d'abord de conserver le texte d'origine et de préciser à quels factoïdes ou relations chaque annotation doit être rattachée.

3.3 Redonner un type aux institutions

Le champ `meta.institutions` est une simple liste de chaînes de caractères, sans indication de nature. Or ces chaînes ne désignent pas des choses de même sorte. Une même

fiche peut porter « Picardie », qui est une des *nations* de l'université, « théologie », qui est une discipline, « Sées », qui est un diocèse, et « Sorbonne », qui est un collège, faisant partie de l'Université de Paris. Le modèle DAPHNÉ les range dans des entités distinctes, ce qui explique le besoin de typage.

Deux traitements s'y emploient, selon que le vocabulaire est fermé ou non. Les nations et les disciplines forment des ensembles finis qu'un médiéviste sait énumérer, et une liste peut suffire à les reconnaître. Deux ont donc été dressées à partir des valeurs effectivement rencontrées dans le corpus, et non d'un référentiel établi. La première rassemble, en 24 entrées, les variantes orthographiques des quatre nations de l'université, France, Picardie, Normandie et nation anglo-allemande. Toute valeur de `meta.institutions` qui s'y trouve devient un GroupP de type « nation ». La seconde recense 23 disciplines, qui deviennent des nœuds Domain, organisés en hiérarchie par des arêtes `DOMAIN_PART_OF`, la Bible relevant de la théologie, la logique des arts, le droit canon du droit. Ces listes couvrent le périmètre traité, mais leur validation par un historien reste un préalable au passage à l'échelle.

Les institutions proprement dites ne se prêtent pas au même traitement. Une liste ne type une valeur que si elle la contient, et il faudrait ici qu'elle en contienne 5 635, dont 3 637 n'apparaissent qu'une seule fois. Ce dernier chiffre suppose d'ailleurs un nettoyage préalable, près de la moitié des valeurs brutes traînant une parenthèse ou une balise de saisie, « France », « France) » et « France$ » comptant sinon pour trois entités différentes.

La queue de cette distribution est faite des localités d'où venaient les étudiants, Nördlingen, Tordesillas, Polebrook ou Beuzeville. Les recenser reviendrait à dresser la carte des lieux d'origine de l'Europe savante, quand il suffit de savoir que ce sont des villes. La section 3.5 évalue un modèle de langage sur cette tâche, typer une chaîne jamais rencontrée à partir de sa forme et de son contexte. Les institutions restent en attendant des GroupP « institution », sans plus de précision.

L'orthographe pose un obstacle d'une autre nature. La même maison apparaît sous « Saint-Jacues » et « Saint-Jacques », la même ville sous « Châlons-en-Champagne » et « Chalons-en-Champagne ». Sans traitement, le graphe compte deux couvents là où il n'y en a qu'un, et les membres du premier ignorent ceux du second.

La distance de Levenshtein compte le nombre minimal de modifications élémentaires, insertion, suppression ou remplacement d'un caractère, nécessaires pour passer d'une chaîne à l'autre. « Augutins » devient « Augustins » en insérant un *s*, soit une distance d'un, et « Sens » devient « Senlis » en insérant deux lettres, soit une distance de deux. Les deux cas n'ont pourtant rien de commun, « Augutins » et « Augustins » sont deux graphies du même mot, quand Sens et Senlis sont deux villes distinctes, à 160 kilomètres l'une de l'autre. Calculée ici sur les formes minusculées et privées de tirets, elle sert donc à trier les candidats plutôt qu'à décider seule.

Les 9 paires identiques après normalisation sont fusionnées directement. Les 52 paires à distance un sont examinées une par une, ce qui est faisable à cette échelle et nécessaire, puisqu'on y trouve aussi bien des coquilles à corriger que des lieux bel et bien distincts. Les 355 paires à distance deux sont écartées en bloc, la proximité graphique n'y signifiant presque jamais l'identité. Les variantes retenues sont enfin regroupées, ce qui permet aux

chaînes de se rejoindre, cinq formes de Châlons convergeant ainsi vers une seule. Les 40 fusions obtenues ramènent les GroupP de 1 124 à 1 084.

La méthode a une limite de principe. Elle ne rapproche que ce qui s'écrit presque pareil, et ne détecte pas les dénominations différentes d'un même lieu. « Châlons-sur-Marne » et « Châlons-en-Champagne » désignent la même ville à sept modifications de distance, et n'ont pu être réunies que par un alias saisi à la main.

Pour les personnes, le rapprochement repose uniquement sur le nom. Une personne est identifiée par la chaîne de son nom, et un parent ou un maître cité dans la notice d'un autre reçoit un nœud à part sous la forme où il est cité. Sur l'export complet, 6 711 personnes n'ont ainsi pas de notice propre, et 604 d'entre elles portent, à la casse ou aux accents près, le nom d'une personne qui en a une. Ces 604 cas constituent des rapprochements possibles avec des personnes disposant d'une notice. Leur identité reste à vérifier à partir d'autres informations, notamment les dates et les relations mentionnées. Aucune fusion n'est faite non plus sur les lieux, les œuvres et les sources, identifiés par leur chaîne nettoyée. Les alias ne s'appliquent qu'aux groupes et aux zones.

Le pipeline a été exécuté sur un sous-échantillon de 3 273 fiches, les lettres A et B, soit environ 15 % du corpus. Sur ce sous-échantillon, le graphe compte environ 4 300 personnes physiques, dont une part n'est connue que par une mention dans la notice d'un autre, 7 736 nœuds de noms, 21 908 factoïdes et 116 172 arêtes, ainsi que 1 084 GroupP après déduplication.

3.4 Mesures d'évaluation

Tous les résultats de classification du rapport, à partir de la section suivante, s'expriment avec les mesures définies ici.

L'exactitude (*accuracy*) est la plus simple, la part des cas où le système répond juste,

$$\mathit{accuracy} = \frac{\text{réponses justes}}{\text{nombre de cas}}.$$

Cette mesure est simple à interpréter, mais elle avantage la classe majoritaire. Si neuf phrases sur dix d'un corpus sont certaines, un système qui répond « certain » à toutes les phrases, sans rien détecter, atteint déjà 90 % d'*accuracy*.

Les trois mesures suivantes se placent du point de vue d'une classe à repérer, l'incertitude par exemple. La précision est la part des alertes qui sont justes, le rappel la part des cas à trouver qui l'ont été, et la spécificité la part des cas négatifs correctement écartés,

$$\text{précision} = \frac{VP}{VP + FP}, \qquad \text{rappel} = \frac{VP}{VP + FN}, \qquad \text{spécificité} = \frac{VN}{VN + FP},$$

où VP, FP, FN et VN comptent les vrais positifs, faux positifs, faux négatifs et vrais négatifs. Précision et rappel varient en sens inverse, une liste de mots élargie trouvant plus et se trompant plus. Le F1 les résume par leur moyenne harmonique,

$$\mathrm{F1} = 2\,\frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}},$$

qui n'est élevé que si les deux le sont.

Avec plus de deux classes, le F1 se calcule d'abord classe par classe, puis les valeurs obtenues sont moyennées, et le choix de la moyenne change ce que la mesure récompense. Le F1 pondéré attribue à chaque classe un poids proportionnel à son effectif. Un système peut donc y obtenir un bon score en ne traitant correctement que les classes fréquentes, car une classe rare mal traitée ne pèse que peu dans la somme. Le F1 macro attribue au contraire le même poids à toutes les classes, quelle que soit leur taille. Échouer sur une classe de dix cas y coûte autant qu'échouer sur une classe de mille. Le F1 macro mesure donc la qualité d'un système sur les classes rares. Formellement,

$$\mathrm{F1}_{\text{pondéré}} = \sum_{c=1}^{C} \frac{n_c}{N} \mathrm{F1}_c, \qquad \mathrm{F1}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c,$$

où $\mathrm{F1}_c$ désigne le F1 de la classe c parmi les C classes, n_c l'effectif de cette classe et N le nombre total de cas.

Ces deux moyennes supposent qu'une réponse désigne une classe et une seule. La classification des institutions de la section 3.5 n'est pas dans ce cas, son jeu de référence admettant plusieurs types pour une même entité, « Paris » valant à la fois pour une ville, un diocèse et une université. La moyenne n'y porte donc ni sur les classes ni sur les étiquettes prises globalement, mais sur les entités. Pour une entité e, soit T_e l'ensemble des types prédits et G_e l'ensemble des types attendus,

$$\text{précision}_e = \frac{|T_e \cap G_e|}{|T_e|}, \qquad \text{rappel}_e = \frac{|T_e \cap G_e|}{|G_e|}, \qquad \mathrm{F1}_e = 2\,\frac{\text{précision}_e \times \text{rappel}_e}{\text{précision}_e + \text{rappel}_e},$$

et le score rapporté est la moyenne non pondérée de ces $\mathrm{F1}_e$ sur les E entités évaluées,

$$\mathrm{F1}_{\text{entité}} = \frac{1}{E} \sum_{e=1}^{E} \mathrm{F1}_e.$$

Chaque entité pèse ainsi le même poids, qu'elle admette un type ou trois, et une réponse partiellement juste garde une part de ses points, là où exiger l'égalité des deux ensembles la compterait fausse.

Toutes ces mesures traitent les classes comme des étiquettes sans ordre. Prédire 4 quand la bonne réponse est 5 y coûte exactement autant que prédire 1, alors que les niveaux de certitude du chapitre 4 sont gradués et que la seconde erreur est bien plus lourde. L'erreur absolue moyenne tient compte de cet ordre, en moyennant l'écart entre le niveau prédit et le niveau attendu,

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|,$$

où y_i est le niveau attendu pour le cas i et $\hat{y}_i$ le niveau prédit. Elle se lit directement en crans d'échelle. Un système qui se tromperait d'un cran sur une phrase sur dix et serait juste partout ailleurs obtiendrait 0,10, et celui qui se tromperait de quatre crans sur la même proportion, 0,40. Contrairement aux précédentes, cette mesure est donc d'autant meilleure qu'elle est basse.

3.5 Typer les institutions par modèle de langage

Des milliers de valeurs ne relèvent d'aucune liste, diocèses, ordres religieux, collèges, universités, abbayes, chapitres, villes et régions confondus. Pour leur redonner un type, un jeu de référence de 150 entités a été construit et annoté manuellement en 9 catégories déterminées empiriquement (diocese, religious_order, college, university, abbey, chapter, city, region, other). Ce jeu tient lieu de vérité pour l'évaluation des modèles. Certains libellés admettent plusieurs types, « Paris » renvoyant à la ville, au diocèse ou à l'université, et le jeu de référence en retient plusieurs pour 18 entités. Onze exemples sont réservés au *few-shot* du prompt, ce qui laisse 139 entités testables.

Les modèles sont évalués via l'API d'Ollama, en version cloud pour les plus gros et à température nulle.

Modèle	**F1**	**AnyOK**
qwen3.5	**87,6 %**	**92,8 %**
devstral	83,3 %	88,5 %
deepseek-v3.1	78,8 %	85,6 %
gpt-oss-120b	77,8 %	88,5 %
gpt-oss-20b	76,7 %	81,3 %
Vote majoritaire des cinq	86,4 %	92,1 %

Table 3.1 – Classification des institutions sur 139 entités. La colonne F1 donne le F1 moyen par entité. AnyOK indique qu'au moins un type prédit est correct. La dernière ligne agrège les cinq modèles par vote, un type étant retenu quand au moins trois le prédisent.

Le meilleur modèle est qwen3.5. Sa confiance déclarée est aussi la plus fiable, et cette qualité n'est pas son privilège. Chaque modèle déclare avec sa réponse un niveau de confiance, et le tableau 3.2 montre que tous les cinq séparent nettement leurs réponses sûres de leurs réponses hésitantes.

	Confiance haute		**Confiance moyenne**		**Confiance basse**	
Modèle	**Part**	**Exactes**	**Part**	**Exactes**	**Part**	**Exactes**
qwen3.5	89,2 %	**83,9 %**	10,1 %	35,7 %	0,7 %	0,0 %
devstral	86,3 %	80,0 %	12,9 %	33,3 %	0,7 %	0,0 %
deepseek-v3.1	78,4 %	77,1 %	20,1 %	25,0 %	1,4 %	50,0 %
gpt-oss-120b	80,6 %	69,6 %	19,4 %	3,7 %	0,0 %	—
gpt-oss-20b	95,4 %	61,3 %	4,6 %	16,7 %	0,0 %	—

Table 3.2 – Répartition des réponses par palier de confiance et justesse dans chacun.

La colonne *part* du tableau dit à quelle fréquence un modèle choisit ce niveau de confiance,

la colonne *exactes* dit à quelle fréquence il donne alors exactement le bon ensemble de types, sans en oublier ni en ajouter.

Le tableau compare la confiance déclarée par chaque modèle à la justesse observée de ses réponses. Pour les cinq modèles, les réponses déclarées avec une confiance haute sont plus souvent exactes que celles déclarées avec une confiance moyenne. Cette association rend l'étiquette utile pour ordonner les réponses, mais ne suffit pas à établir que la confiance est calibrée. gpt-oss-20b se déclare sûr plus souvent que qwen3.5, 95,4 contre 89,2 %, et se trompe pourtant vingt-trois points plus souvent dans ces cas-là. Une calibration complète demanderait que chaque palier corresponde à une probabilité de réussite définie et vérifiée sur un nombre suffisant d'exemples.

Ces prédictions pourraient servir de propositions à valider par une relecture humaine experte. Le désaccord entre modèles et leur confiance déclarée permettraient de prioriser cette relecture, complétée par un contrôle aléatoire des réponses consensuelles.

3.6 Bilan du chapitre

Ce chapitre devait convertir la part régulière des fiches en un graphe interrogeable. Pour le périmètre traité, le graphe compte environ 4 300 personnes et 116 000 arêtes. Chaque factoïde garde sa rubrique d'origine, et les dates approximatives y survivent avec leur nuance. Les nations et les disciplines ont retrouvé un type par listes. Pour les autres institutions, qwen3.5 atteint 87,6 % de F1 sur le jeu de test, avec une confiance déclarée assez fiable pour cibler la relecture humaine, mais ses prédictions n'ont pas été versées dans le graphe.

Ce graphe est un premier état. Sa provenance s'arrête à la notice, faute de lire les références de chaque bloc. Les rapprochements de personnes qu'il opère par le nom restent à vérifier, et les fusions orthographiques ne concernent que les groupes. Les annotations du texte libre restent à y intégrer, ce qui suppose de conserver le texte d'origine des factoïdes. Les listes attendent la validation d'un historien, et la qualité de la conversion n'a pas été mesurée.

Les règles ont ainsi converti les champs structurés et conservé les réserves que la saisie rendait déjà explicites, les dates approximatives et les « ? » accolés à une valeur structurée. Les réserves que les rédacteurs formulent dans le texte libre, le « peut-être », le « vers », le « douteux », restent à analyser et à intégrer. Le chapitre 4 le fait par une première tâche, attribuer à chaque phrase un degré de certitude selon un barème. Cette tâche se prête à une comparaison entre des méthodes entraînées sur le corpus et des modèles de langage qui reçoivent les règles d'annotation dans leur consigne. Le chapitre 5 introduit ensuite une annotation qui distingue, dans une même phrase, les informations concernées.

Chapitre 4

Mesurer l'incertitude

Il n'est pas absolument certain que le boursier du Collège de Dormans-Beauvais soit la même personne que l'évêque de Paris : Thierry Kouamé considère toutefois que cela est très probable ;

Studium Parisiense, fiche n° 19155, Dionysius de Molendino

Ce chapitre prolonge un travail engagé avant le stage pour un article en préparation dans la revue *Data & Knowledge Engineering*. La tâche prend une phrase de notice en entrée et rend un score de certitude de 1 à 5, attribué selon un barème. Plus de deux mille phrases annotées permettent d'entraîner des méthodes supervisées et de poser une question : ce score est-il mieux prédit par des méthodes entraînées sur le corpus ou par des modèles de langage généralistes qui reçoivent le barème dans leur consigne ?

Formellement, une phrase x reçoit un niveau de référence $y(x) \in \{1, \ldots, 5\}$, fixé par le barème de la section 4.3, où 5 note un fait affirmé sans réserve et 1 un fait que le rédacteur tient pour faux. Une méthode est une fonction qui associe à x un niveau prédit $\hat{y}(x)$, sans accès au reste de la notice. Les méthodes sont comparées sur les 487 phrases du jeu de test par les mesures de la section 3.4, l'exactitude, qui mesure la proportion de phrases recevant le niveau attendu, le F1 macro, qui donne le même poids aux cinq niveaux, et l'erreur absolue moyenne, qui mesure l'écart en crans.

4.1 Quatre méthodes pour mesurer le doute

L'article, engagé par l'équipe encadrante, prolonge sa première détection d'incertitude sur ces données [40]. Il quantifie l'incertitude de phrases de Studium en comparant un SVM, un modèle CamemBERT fine-tuné [48], une méthode par dictionnaire et une architecture en deux étages (coarse2fine). La tâche, l'échelle et ces quatre méthodes sont acquises. La contribution du stage commence après. Elle audite puis réannote le jeu de phrases qui sert aux évaluations, réimplémente les quatre méthodes et y ajoute neuf modèles de langage

évalués sans entraînement. L'objectif est de conduire l'article à la soumission avec ces compléments.

Ces quatre méthodes reposent sur des compromis différents, entre les ressources qu'elles exigent et ce qu'elles voient de la phrase, et leur ordre de présentation ne vaut pas classement. La méthode par **dictionnaire** cherche dans la phrase les entrées d'une liste de mots, chacune portant un score de 1 à 5. Le dictionnaire lui-même vient de la première étude [40], où il servait à détecter l'incertitude sans la graduer. L'article y ajoute le score de chaque entrée, fixé par un panel d'experts, puis l'agrégation de ces scores à l'échelle de la phrase. Les auteurs ont construit ce dictionnaire en deux temps. Une liste de départ de 65 termes a d'abord été réunie depuis des ressources linguistiques françaises et par traduction de la liste anglaise de Chen et al. [14], répartie en adjectifs, adverbes, noms, verbes et expressions. Cette liste a ensuite été étendue par synonymie. WordNet est une base lexicale qui regroupe les mots de l'anglais en ensembles de synonymes, les *synsets*, reliés entre eux par des relations sémantiques, et WoNeF en est une traduction française construite automatiquement. Les synonymes que WoNeF donne pour chaque terme de la liste ont été ajoutés, l'opération répétée sur chaque mot ajouté jusqu'à ce que plus aucun nouveau terme n'apparaisse, et le résultat élagué à la main. Le dictionnaire final compte 190 entrées. Le tableau 4.1 en donne un aperçu.

Catégorie	**Exemples**	**Départ**	**Final**
Adjectifs	douteux, incertain, probable, vraisemblable, possible, présumé	11	57
Adverbes	approximativement, grossièrement, environ, probablement, presque	6	24
Noms	incertitude, probabilité, estimation, doute, hasard, possibilité	9	29
Verbes	penser, croire, douter, hésiter, paraître, sembler, estimer	9	50
Expressions	plus ou moins, aux environs, à peu près, l'hypothèse que, autour de	30	30
Total		**65**	**190**

Table 4.1 – Le dictionnaire de l'article, avant et après extension par les synonymes de WoNeF.

Scorer une phrase revient alors à y chercher les entrées du dictionnaire et à agréger les scores trouvés, une phrase sans aucune entrée valant 5. La méthode hérite des limites d'une liste, car un mot absent n'est pas vu et un mot présent compte toujours pareil.

Le **SVM** (*support vector machine*, littéralement « machine à vecteurs de support ») apprend au lieu de suivre une liste. Chaque phrase est d'abord transformée en un vecteur de caractéristiques. Chaque phrase devient donc un point dans un espace à plusieurs milliers de dimensions. L'entraînement consiste à lui montrer des phrases dont le niveau est déjà

annoté, et il cherche la frontière qui sépare les niveaux en laissant de part et d'autre la marge la plus large possible. Il découvre ainsi des régularités qu'aucune liste n'aurait énoncées, mais il ne voit de la phrase que ce que ces caractéristiques encodent.

CamemBERT utilise des représentations contextuelles apprises, ce qui réduit le recours à des caractéristiques choisies manuellement. C'est un modèle de langage pré-entraîné sur de grandes quantités de français, qui représente chaque mot en fonction de son contexte, de sorte que le *ca.* d'un titre d'ouvrage et celui d'une datation ne sont pas encodés de la même façon. Ce modèle ne produit toutefois que des vecteurs. Pour qu'il classe, une tête de classification lui est ajoutée, deux couches linéaires qui lisent le vecteur résumant la phrase et le ramènent à l'un des cinq niveaux. Cette tête part de poids aléatoires. Son *fine-tuning* se poursuit sur les phrases annotées du corpus pour qu'il produise en sortie l'un des cinq niveaux. Il n'y a plus de caractéristiques à choisir à la main, mais il faut des exemples annotés, et le modèle obtenu ne sait faire que cette tâche.

L'architecture **en deux étages** part d'un constat sur les données plutôt que sur les modèles. Le corpus est massivement certain, et un classifieur qui traite les cinq niveaux d'un coup consacre l'essentiel de sa capacité à la classe dominante. La décision est donc coupée en deux. Un premier CamemBERT ne tranche qu'une question binaire : la phrase est-elle certaine ou non. Celles qu'il juge certaines reçoivent directement le niveau 5. Les autres passent dans un second CamemBERT, entraîné sur les seules phrases incertaines, qui les répartit entre les niveaux 1 à 4. Chacun des deux modèles traite ainsi un problème plus simple que le problème d'origine.

4.2 Autopsie d'un corpus

Les données de ces expériences sont des phrases extraites des fiches de Studium, découpées à l'unité de la phrase et annotées à la main d'un score de 1 à 5.

Le premier travail a été une étude de ces données qui a mis en évidence plusieurs problèmes. Ainsi, sur les 2 432 annotations du jeu complet, le niveau 1 de l'échelle n'était jamais utilisé. On y comptait 38 paires de phrases dupliquées portant des scores divergents, et 67 phrases figuraient à la fois dans le jeu d'entraînement et dans le jeu de test, dont 22 avec des scores différents. Une partie du corpus souffrait de problèmes d'encodage, 546 lignes touchées dont 329 ont pu être réparées automatiquement, le reste étant irrécupérable. L'analyse par marqueur a révélé enfin des incohérences internes, comme « vers + année » qui recevait selon les fiches des scores de 2 à 5, alors que le barème de l'article en prescrivait 3.

Une fois les doublons retirés et les phrases partagées ramenées au seul jeu de test, il reste 1 747 phrases d'entraînement et 487 phrases de test. Les 2 234 phrases des deux jeux ont donc été réannotées avant toute nouvelle comparaison.

4.3 Le barème d'annotation

Le barème part de ce que l'échelle mesure. Pour être tout à fait clair, le score note la façon dont l'historien qui a rédigé la notice s'engage sur ce qu'il écrit. « Il est né à Paris » vaut 5 parce que la phrase l'affirme sans réserve, quand bien même un autre document la contredirait. « Il est probablement né à Paris » vaut 4 parce que l'historien a choisi d'écrire « probablement », quand bien même la chose serait acquise par ailleurs.

Les niveaux de l'échelle 5, 4 et 3 vont tous dans le sens du fait tenu pour vrai, avec une force qui décroît, affirmé, probable, possible. Les deux derniers ne vont plus dans ce sens. Au niveau 2, l'auteur penche vers le faux ou déclare qu'il ne sait pas. Au niveau 1, il tient le fait pour faux, ce qui vaut par exemple pour une œuvre reconnue pseudépigraphe, c'est-à-dire attribuée à tort à un auteur qui ne l'a pas écrite.

Les marqueurs sont classés en sept familles, détaillées avec le score de chacun dans le tableau 4.2.

Famille	**Marqueurs typiques**	**Score**
Aucun marqueur	« Il est né à Paris »	5
A. Valeur approximative	« vers 1300 », « entre 1250 et 1255 », « oncle ou grand-oncle »	3
	« né près de Paris », le lieu exact étant inconnu	4
B. L'auteur nuance	« semble », « probablement », « sans doute », « vraisemblablement »	4
	« peut-être », « il serait né à Reims », « il n'est pas impossible que »	3
C. L'auteur cite un tiers	« selon X », « d'après X », « on dit que », « attesté par »	4
	« on a *aussi* suggéré que », une autre version existant	3
D. Attribution d'œuvre	« attribué à X », « traditionnellement attribué à X »	4
	« attribué à X ou à Y », « attribué par certains »	3
	« attribution douteuse », « attribution contestée »	2
	« faussement attribué », « pseudépigraphe »	1
E. Point d'interrogation	« Bachelier en théologie (Paris) ? »	3
F. Doute déclaré	« incertain »	3
	« n'est pas assuré », « douteux », « mis en doute »	2
	« très douteux », « apocryphe », « authenticité rejetée »	1
G. Ignorance déclarée	« on ignore », « date inconnue », « date conjecturale »	2

Table 4.2 – Les familles de marqueurs et le score qu'elles appellent. Une attribution ne vaut jamais 5, puisqu'elle signale par construction un savoir de seconde main.

Une règle encadre l'attribution du score. Quand plusieurs marqueurs coexistent, le score de la phrase est le minimum des leurs.

Entre les deux annotations, 307 phrases changent de score, soit 13,7 %. L'écart est d'un cran pour 231 d'entre elles, de deux crans pour 63, et de trois ou quatre pour 13 seulement. Le taux diffère un peu selon le jeu, 16,2 % sur le test contre 13,2 % sur l'entraînement. Le mouvement est enfin orienté, puisque 238 phrases descendent vers plus d'incertitude contre 69 qui remontent.

4.4 Protocole d'évaluation

Neuf modèles de langage généralistes, utilisés par prompt, ont été évalués aux côtés des quatre méthodes de référence, choisis dans sept familles différentes pour éviter qu'un résultat ne tienne aux particularités d'un seul constructeur. On y trouve nemotron-3-ultra, minimax-m3, glm-5.2, claude-sonnet-5, gemma4-31b, deepseek-v4-pro et sa variante allégée, gpt-oss-120b et gpt-oss-20b, ce dernier étant vingt fois plus petit que les plus gros du lot[1]. Aucun n'a été entraîné sur le corpus. Tous reçoivent la même consigne, reproduite en annexe A, traitent les phrases une par une et doivent citer le marqueur sur lequel ils s'appuient avant de donner leur score. Interrogés par API, ils ne recourent à aucune autre information que ce qui leur est fourni dans le prompt. En particulier, ils n'ont pas accès à un historique des discussions avec un utilisateur.

Une dernière précaution tient à la taille du corpus. Sur 487 phrases de test, un écart d'un point d'*accuracy* ne représente que cinq phrases, et rien dans le chiffre seul ne dit si ces cinq phrases viennent de la méthode ou du hasard de l'échantillon. Les modèles CamemBERT sont donc entraînés avec trois graines différentes[2], ce qui permet de rapporter la moyenne et l'écart-type.

Chaque écart entre deux systèmes est en outre accompagné d'un test, qui répond toujours à la même question, cet avantage tiendrait-il sur d'autres phrases que celles-ci ? Le test de McNemar[3] compare les deux systèmes phrase par phrase et ne retient que celles où ils divergent, l'un ayant juste et l'autre faux. Si ces désaccords se répartissent à peu près

1. Huit des LLM ont été servis par Ollama en version cloud, sous les identifiants `nemotron-3-ultra:cloud` (https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16), `minimax-m3:cloud` (https://huggingface.co/MiniMaxAI/MiniMax-M3), `glm-5.2:cloud` (https://huggingface.co/zai-org/GLM-5.2), `gemma4:31b-cloud` (https://huggingface.co/google/gemma-4-31B-it), `deepseek-v4-pro:cloud` (https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro), `deepseek-v4-flash:cloud` (https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash), `gpt-oss:120b-cloud` (https://huggingface.co/openai/gpt-oss-120b) et `gpt-oss:20b-cloud` (https://huggingface.co/openai/gpt-oss-20b). claude-sonnet-5 a été interrogé directement par l'API Anthropic (https://www.anthropic.com/news/claude-sonnet-5).

2. La graine (*seed*) fixe la part d'aléa de l'entraînement, l'ordre de présentation des exemples et l'initialisation des poids. Deux entraînements à graines différentes donnent deux modèles légèrement différents, d'où la moyenne sur trois.

3. https://fr.wikipedia.org/wiki/Test_de_McNemar.

également dans les deux sens, l'avance affichée par le total ne repose sur rien de solide. Le F1 macro ne se prête pas à ce décompte, puisqu'il s'obtient en moyennant des classes et non en additionnant des phrases correctes. On lui applique donc un bootstrap apparié [4]. L'idée est de simuler d'autres jeux de test possibles à partir du seul dont on dispose. Un tirage consiste à prendre au hasard, avec remise, 487 phrases parmi les 487 du test, si bien que certaines phrases y figurent plusieurs fois et d'autres pas du tout. Sur chacun de ces jeux simulés, des milliers au total, le F1 macro des deux systèmes est recalculé, toujours sur les mêmes tirages pour l'un et pour l'autre. Si le classement des deux systèmes s'inverse dans une part notable des tirages, l'écart observé sur le vrai test peut n'être qu'un accident d'échantillonnage.

Sans ces tests, le tableau des résultats se lirait comme un classement, alors qu'une bonne part des écarts qu'il affiche ne se distinguent pas du bruit.

4.5 Résultats

Le tableau 4.3 rejoue toutes les méthodes sur le couple cohérent formé par le corpus réannoté, et y ajoute neuf modèles de langage évalués sans aucun entraînement.

Méthode	Famille	Acc.	Acc. ±1	F1 pond.	F1 macro	MAE
Modèles généralistes, utilisés par consigne						
nemotron-3-ultra	NVIDIA	**96,3**	98,8	**96,4**	82,4	0,051
minimax-m3	MiniMax	96,1	99,0	96,0	83,9	0,051
glm-5.2	Zhipu	95,9	99,0	95,9	81,3	0,053
claude-sonnet-5	Anthropic	95,7	98,6	95,6	**87,1**	0,060
gemma4-31b	Google	95,7	**99,4**	95,6	86,7	**0,049**
deepseek-v4-pro	DeepSeek	95,3	98,4	95,5	77,8	0,068
gpt-oss-120b	OpenAI	95,3	98,8	95,0	73,2	0,060
deepseek-v4-flash	DeepSeek	94,9	98,8	94,7	79,9	0,064
gpt-oss-20b	OpenAI	91,8	97,9	91,4	67,9	0,107
Méthodes de référence de l'article						
CamemBERT fine-tuné	—	94,6 ±0,4	98,8	94,4 ±0,4	71,2 ±0,5	0,076
coarse2fine (deux étages)	—	92,1 ±0,3	99,0	92,0 ±0,3	67,0 ±0,1	0,086
SVM	—	87,3	95,7	86,7	72,6	—
Dictionnaire	—	75,8	85,2	71,2	38,9	0,421

Table 4.3 – Toutes les méthodes sur le test réannoté (n=487). La colonne Acc. ±1 compte aussi juste un score à un niveau du gold, nom usuel de l'annotation de référence tenue pour vraie, et la MAE, définie à la section 3.4, donne l'écart moyen au gold en crans d'échelle.

La colonne ±1 situe la gravité des erreurs. En tolérant un cran d'écart, les neuf modèles

4. https://fr.wikipedia.org/wiki/Bootstrap_(statistiques).

de langage dépassent 97,9 %, le SVM atteint 95,7 %, et les prédictions à deux niveaux ou plus du gold restent rares, moins de 2 % pour huit modèles sur neuf. Le dictionnaire fait moins bien, 14,8 % du test tombant à deux crans ou plus, son défaut à 5 projetant loin toute phrase dont le marqueur lui échappe.

La MAE résume la même idée en un seul nombre. Elle moyenne l'écart entre le niveau prédit et le niveau attendu, et se lit en crans d'échelle. Toutes les méthodes pour lesquelles elle a pu être calculée, le dictionnaire excepté, restent sous 0,11, autrement dit à moins d'un neuvième de cran en moyenne, et les huit meilleurs modèles de langage sous 0,07. Le dictionnaire est à 0,421, plus de cinq fois plus loin que CamemBERT. Cette mesure complète l'exactitude en tenant compte de l'amplitude des erreurs. Sur l'*accuracy* il conservait 75,8 %, un chiffre qui pouvait passer pour honorable, alors qu'il se trompe non seulement plus souvent mais beaucoup plus largement.

Cette mesure déplace aussi le classement de tête. gemma4-31b, qui n'arrive qu'en cinquième position sur l'*accuracy*, obtient la MAE la plus basse et la meilleure tolérance à un cran, quand nemotron-3-ultra mène sur l'*accuracy* et claude-sonnet-5 sur le F1 macro. Trois mesures désignent ainsi trois vainqueurs différents. Les écarts de scores sont relativement faibles, et il est donc difficile d'affirmer sans nuance qu'un modèle est plus adapté qu'un autre sur cette tâche.

Comparaison avec CamemBERT	***Accuracy***	**F1 macro**
nemotron-3-ultra	+2,7 ($p = 0{,}04$)	+12,5 ($p = 0{,}003$)
minimax-m3	+2,5 (n.s.)	+14,0 ($p = 0{,}01$)
claude-sonnet-5	+2,1 (n.s.)	+17,2 ($p = 0{,}01$)
gemma4-31b	+2,1 (n.s.)	+16,9 ($p = 0{,}03$)
glm-5.2	+2,3 (n.s.)	+11,4 ($p = 0{,}03$)
deepseek-v4-pro	+1,6 (n.s.)	+7,9 (n.s.)
gpt-oss-120b	+1,6 (n.s.)	+3,3 (n.s.)
deepseek-v4-flash	+1,2 (n.s.)	+10,1 (n.s.)
gpt-oss-20b	−1,8 (n.s.)	−2,0 (n.s.)

Table 4.4 – Écarts à CamemBERT et leur significativité (McNemar exact pour l'*accuracy*, bootstrap apparié pour le F1 macro. « n.s. » vaut pour $p > 0{,}05$). L'avantage en *accuracy* ne résiste au test que pour nemotron-3-ultra.

Les modèles généralistes utilisés par consigne égalent l'encodeur adapté par apprentissage supervisé et le dépassent sur les classes rares. Huit sur neuf affichent une *accuracy* supérieure à celle de CamemBERT, mais le test de McNemar ne rejette l'égalité que pour un seul d'entre eux (tableau 4.4). En *accuracy*, les deux familles sont donc comparables, sans que l'une devance l'autre.

L'*accuracy* est cependant saturée par la classe 5, qui représente 62 % du test et que tout le monde traite bien. Sur le F1 macro, qui pondère également toutes les classes, la différence atteint 11 à 17 points et résiste au test pour les cinq meilleurs modèles.

L'explication tient à la nature des deux approches. CamemBERT doit induire la règle à partir d'exemples, dont la classe 1 n'a que 15 occurrences, là où un modèle généraliste reçoit la règle énoncée dans son prompt, ce qui le rend plus performant sur les classes rares.

4.6 L'accord entre modèles

Les scores agrégés ne disent pas si les neuf modèles réussissent aux mêmes endroits. Le tableau 4.5 compte, pour chacune des 487 phrases du test, combien des neuf modèles retrouvent le niveau exact du gold.

Modèles corrects	Phrases	Part
9 sur 9	423	86,9 %
8	26	5,3 %
7	8	1,6 %
6	6	1,2 %
5	5	1,0 %
4	5	1,0 %
3	3	0,6 %
2	4	0,8 %
1	4	0,8 %
0 sur 9	3	0,6 %

Table 4.5 – Nombre de modèles donnant le niveau exact du gold, par phrase (487 phrases, neuf modèles).

Sur 423 phrases, les neuf y parviennent ensemble. Les 210 erreurs restantes, soit 4,8 % des 4 383 prédictions, se concentrent sur 64 phrases, et les dix-neuf phrases où au plus quatre modèles retrouvent le gold sont exactement les dix-neuf où le vote majoritaire le contredit. L'erreur des modèles de langage se cristallise sur un petit noyau de phrases.

Trois phrases échappent aux neuf modèles à la fois, deux attributions d'œuvre et un savoir rapporté. « Oeuvre attribuée par certains biographes dont il n'est pas fait mention dans le nécrologe du couvent dominicain de Sienne » reçoit le score de 3 de la part de huit modèles, le score que le barème donne à « attribué par certains », quand le gold descend à 2. Le désaccord porte ici sur la prise en compte de l'absence au nécrologe comme contre-évidence.

4.7 Bilan du chapitre

La contribution du chapitre tient en plusieurs points. Un jeu de référence assaini d'abord, 2 234 phrases réannotées suivant un barème explicite, là où le jeu hérité mêlait doublons, fuites et scores contradictoires. Les quatre méthodes de l'article ensuite, réimplémentées et

réentraînées, ce qui inverse au passage la hiérarchie attendue, le CamemBERT direct dépassant l'architecture en deux étages. Neuf modèles de langage enfin, évalués sans aucun entraînement, qui égalent le modèle affiné en exactitude et le dépassent nettement sur les niveaux peu représentés du corpus.

Pour le graphe, ce score a une limite. Même correctement prédit, il qualifie la phrase entière et ne dit pas quelle information y est incertaine. Or une phrase porte souvent plusieurs informations, auxquelles le rédacteur n'accorde pas le même crédit comme , un lieu mis en doute à côté d'une date seulement approchée et d'un grade affirmé sans réserve. Un score unique ne restitue pas ces distinctions et ne permet pas de déterminer, à lui seul, quel factoïde ou quelle relation du graphe doit recevoir l'annotation. Le chapitre suivant étudie une annotation qui le formalise.

Chapitre 5

Détecter et qualifier l'incertitude dans Studium

?Étudiant : il est réputé avoir étudié à Paris sans doute vers 1150-1160, mais il est difficile de trouver les sources de cette affirmation ?

Studium Parisiense, fiche n° 58824, Henricus Bretislav

Le score du chapitre précédent qualifie une phrase entière. Pour poser une réserve sur le graphe de connaissances que l'on souhaite obtenir, il faut savoir en plus quelle information de la phrase elle concerne, afin de l'attacher au bon factoïde ou à la bonne relation. Ce chapitre étudie une annotation qui identifie les marqueurs, distingue trois phénomènes, l'incertitude, l'imprécision et l'incomplétude, et désigne l'élément concerné. Cette annotation prépare le rattachement des réserves aux factoïdes et aux relations de DAPHNÉ, rattachement qui n'a pas été finalisé pendant le stage. Le jeu de référence et l'échelle sont distincts de ceux du chapitre précédent, dont les scores ne sont pas réutilisés.

Il faut donc savoir trois choses d'une phrase, si elle exprime un doute, sur quoi ce doute porte, et avec quelle force. Une simple recherche de mots répond déjà presque à la première question, mais elle ne dit rien des deux autres, et c'est ce manque qui conduit à un modèle de langage, puis à une définition de l'incertitude propre au corpus.

5.1 Ce qu'une liste de mots repère

Avant de mobiliser un modèle de langage, il fallait mesurer ce qu'une simple recherche de mots donne sur ce corpus. Un jeu de référence de 147 phrases de notices a donc été annoté à la main, 83 incertaines et 64 certaines, sur quatre niveaux, la détection binaire, la portée (le doute touche-t-il l'énoncé entier ou une relation précise ?), le type de relation concernée et l'entité visée.

Beaucoup de ces phrases sont d'une grande simplicité. L'historien écrit « Bachelier en théologie (Paris) ? » ou « Étudiant (Paris) avant 1487 ? », et le point d'interrogation final porte seul le doute, sans un mot pour l'accompagner. Sur les 83 phrases incertaines du jeu, 19 contiennent un « ? » et 9 n'ont rien d'autre. Aucune des 64 phrases certaines n'en porte. Le signe se cherche avec une expression régulière d'un caractère, et il ne se trompe jamais ici.

Les marqueurs lexicaux ont été cherchés avec un dictionnaire construit à partir de la typologie des marqueurs de modalisation de Hyland [33], des marqueurs de BioScope et de ceux des corpus cliniques français CAS et ESSAI [19]. Cette liste de 98 mots traduits de l'anglais, à laquelle s'ajoute le « ? », atteint 87,4 % de F1 sur la détection binaire, avec 97,1 % de précision et 79,5 % de rappel (section 3.4). Autrement dit, elle ne se déclenche presque jamais à tort, mais laisse passer une phrase incertaine sur cinq.

Une bonne part de ce qui lui échappe se corrige sans changer d'approche. La liste contient *sembler*, *douteux* et *pourrait*, là où les notices écrivent « il semble », « douteuse » et « pourraient », et treize des dix-sept phrases manquées tiennent à ce seul écart de forme. Conjuguer les verbes et ajouter les féminins améliore donc les résultats. Mais plus la liste s'allonge, plus elle se déclenche, y compris là où personne ne doute de rien. Ce qu'on gagne en rappel peut donc se perdre en précision.

Deux mots montrent ce qu'un réglage de la liste ne peut pas corriger. Le *ca.*, abréviation du latin *circa*, « environ », d'une datation approximative signale un doute réel, « le traité est donc datable ca. 1300 ». Le même *ca.* recopié dans un titre d'ouvrage n'en signale aucun, « *Kurienuniversität und stadtrömische Universität von ca. 1300 bis 1471* »[1] n'étant qu'une référence bibliographique. La chaîne de caractères est pourtant identique dans les deux phrases. Le *ou* se comporte de la même façon. Il pose deux hypothèses concurrentes dans « la *disputatio* est lue par lui à Rome ou à Anagni », où l'une des deux villes est la bonne sans qu'on sache laquelle, et une simple coordination dans « cet accord interdit au chancelier de réclamer de l'argent ou un serment aux licenciés », où les deux sont interdits. Seul le reste de la phrase décide, ce qu'une recherche de mots ne peut pas voir.

Une autre limite est plus tenace. Même quand la liste se déclenche à bon droit, elle ne dit rien de ce qui est douteux. Dans « Bachelier en théologie (Paris ?) vers 1278 », deux marqueurs cohabitent et ne visent pas la même chose, le « ? » met en doute le lieu du grade et le *vers* en approxime la date, tandis que le grade lui-même n'est pas contesté. Une recherche de mots rend un bit pour toute la phrase, là où un graphe de connaissances a besoin de savoir quelle arête affaiblir.

Elle ne dit pas davantage à quel point la chose est douteuse. « Il est sûrement né à Paris » et « il est douteux qu'il soit né à Paris » déclenchent toutes deux la liste et en ressortent également incertaines, alors que la première penche vers l'affirmation et la seconde vers son contraire. Un historien n'accordera pas le même crédit aux deux, et un graphe qui les traite pareil perd ce que le rédacteur a pris soin d'écrire. La certitude est donc graduée sur cinq niveaux plutôt que signalée par oui ou non (section 5.2).

1. « Université de la Curie et université de la ville de Rome, d'environ 1300 à 1471 », traduction effectuée sur DeepL : https://www.deepl.com/fr/translator (consulté le 01/09/2026).

Un modèle de langage, lui, produit une sortie structurée. On lui donne la phrase accompagnée du nom du champ dont elle provient, et on lui demande un objet JSON plutôt qu'une réponse en français. La figure 5.1 montre ce que qwen3.5 rend sur cette même phrase.

L'objet se lit en deux temps. Le premier niveau répond à la question binaire, `has_uncertainty` vaut vrai et `markers` énumère les chaînes qui l'ont déclenché, ici le « ? » et le « vers ». Une liste de mots atteint déjà ce niveau. Le second, `annotations`, est celui qu'elle n'atteint pas, avec une entrée par incertitude distincte. Le `scope` indique où l'incertitude se loge dans le graphe, la valeur `edge` désignant une relation plutôt qu'un nœud ou un factoïde entier. `edge_type` nomme cette relation dans le vocabulaire DAPHNÉ, `TookPlaceAt` pour le lieu d'un événement et `OccurredAt` pour sa date. `target` et `target_type` désignent l'entité visée et sa catégorie. `uncertain_element`, enfin, reformule en clair ce qui est mis en doute, ce qui permet de contrôler l'annotation sans relire le reste.

La phrase donne ainsi deux entrées et non une, rattachées à deux arêtes différentes. Le grade de bachelier n'apparaît dans aucune, donc il reste certain. Cette information désigne l'arête du graphe où poser l'attribut de certitude, ce qu'une recherche de mots ne fournit pas. Sur l'ensemble des 147 phrases, qwen3.5 atteint 99 % de F1 en détection binaire, 85 % sur la portée, 82 % sur le type de relation et 78 % sur la cible.

```
{ "has_uncertainty": true,
  "markers": [ "?", "vers" ],
  "annotations": [
    { "scope": "edge", "edge_type": "TookPlaceAt",
      "target": "Paris", "target_type": "Place",
      "uncertain_element": "lieu du baccalauréat (Paris)" },
    { "scope": "edge", "edge_type": "OccurredAt",
      "target": "1278",  "target_type": "Time",
      "uncertain_element": "date approximative (vers 1278)" } ] }
```

Figure 5.1 – Sortie de qwen3.5 sur « Bachelier en théologie (Paris ?) vers 1278 ». Le « ? » est rattaché au lieu du grade, le « vers » à sa date, et le grade lui-même reste certain.

5.2 Un modèle d'incertitude pour Studium

Savoir qu'une phrase doute ne suffit pas. Pour savoir *quoi* annoter, il faut un modèle de l'incertitude propre au corpus. Le modèle que nous présentons a été construit à partir d'une lecture manuelle d'une centaine de fiches, puis d'une annotation structurée de 191 éléments répartis sur 97 fiches. Chaque annotation décrit un phénomène, son marqueur, sa portée textuelle, la catégorie d'information visée et, lorsqu'il s'agit d'une incertitude, un niveau de certitude.

Cette étude de corpus a imposé une distinction entre trois phénomènes que le mot « incertitude » écrase habituellement, distinction que la littérature sur l'extraction de connaissances range parmi les imperfections des données, avec l'ambiguïté, le flou et l'incohé-

rence [38] : l'**incertitude** proprement dite (le fait est-il vrai ? « peut-être frère du couvent d'Uppsala »), l'**imprécision** (le fait est assuré mais la valeur est vague : « vers 1241 ») et l'**incomplétude** (la valeur est absente : « Cursus : ? »). Les trois se croisent librement, et le même « ? » relève de l'une ou de l'autre selon le contexte. Ce signe est aussi le plus fréquent, 44 % des éléments annotés en portant un en suffixe, auxquels s'ajoutent les « ? » préfixes et les champs vides. Les marqueurs lexicaux (« probablement », « sans doute ») sont comparativement rares. C'est une spécificité de ce corpus par rapport à d'autres présents dans la littérature.

Le cadrage retenu se formule ainsi. L'unité annotée est un extrait de fiche x, formé de la valeur d'un champ, du nom de ce champ et, s'il existe, du commentaire du rédacteur. Une annotation est un quintuplet $a = (p, m, s, c, \ell)$, et un extrait reçoit un ensemble d'annotations $A(x)$, vide lorsqu'aucun des trois phénomènes n'y est annoté. La figure 5.2 montre un extrait et ses deux annotations attendues.

```
Entrée
  Fiche : MAGNUS Augustini
  Champ Studium : Cursus > Grades
  Valeur : ?Magister (P), avant 1347.
  Commentaire : il est cité une unique fois comme magister en 1351

Sortie attendue
  {"annotations": [
    {"phenomenon": "incertitude", "trigger": "?",
     "scope_text": "Magister (P)", "target_component": "statut",
     "certainty_level": 2, "precision": "NA",
     "missing_type": "NA",
     "description_libre": "L'incertitude porte sur le statut,
        de l'individu, car Magister (P) n'est pas certain."},
    {"phenomenon": "imprécision", "trigger": "avant",
     "scope_text": "avant 1347", "target_component": "date",
     "certainty_level": null, "precision": "imprécis",
     "missing_type": "NA",
     "description_libre": "L'imprécision porte sur la datation,
        avant 1347 donne une borne ou une période non exacte."}
  ] }
```

Figure 5.2 – Un extrait du jeu de référence et son annotation attendue, deux phénomènes dans un même champ, un « ? » sur le statut et un « avant » sur la date. « NA » marque une clé sans objet pour le phénomène annoté.

Le phénomène p, clé `phenomenon`, prend l'une des trois valeurs incertitude, imprécision et incomplétude. Le marqueur m, clé `trigger`, est le mot ou le signe de x qui le signale, et la portée s, clé `scope_text`, le segment de x que m affecte, vide quand le champ entier est visé. La catégorie c, clé `target_component`, indique le type d'information concerné, par exemple une date ou un lieu, parmi huit valeurs, lieu, date, statut, institution, identité, relation, action et champ. Le choix de la relation DAPHNÉ à annoter, que désignait le format de la section 5.1, est laissé à une étape ultérieure. Le niveau

$\ell \in \{1, \ldots, 5\}$, clé `certainty_level`, gradue la certitude et n'est renseigné que pour l'incertitude. Dans l'exemple, « avant 1347 » reçoit une seconde annotation, d'imprécision, dont la catégorie est la date et qui n'a pas de niveau, ce qui montre pourquoi un extrait peut recevoir plusieurs annotations. Pour l'incomplétude, la clé `missing_type` distingue le manque total du manque partiel. L'échelle de ℓ demande de décider où passent les frontières entre les degrés. Le tableau 5.1 donne l'échelle telle qu'elle a été arrêtée, avec pour chaque niveau ce que le rédacteur de la notice signale et un exemple du corpus.

Niv.	Ce que le rédacteur signale	Exemple du corpus
5	**Certain.** Aucun marqueur de doute	« Licencié en théologie (Paris), 1504 (19 février) »
4	**Très probable.** Forte confiance sans certitude, avec « sans doute », « très probablement », « certainement »	« Son frère est sans doute HEMMING Gregersson »
3	**Probable.** Information plausible, engagement plus faible, avec « probable », « il semble »	« le style de ses sermons semble dénoter un universitaire »
2	**Incertain.** Possibilité parmi d'autres, avec « peut-être », « il est possible que », et le « ? »	« Son parent est peut-être Guillelmus de VISSACO »
1	**Très incertain.** Information donnée pour douteuse ou contestable	« l'attribution est douteuse »

Table 5.1 – Les cinq degrés de certitude, du plus assuré au plus fragile.

L'échelle se lit du plus assuré au plus fragile. Le niveau 5 est le cas par défaut, celui d'une information posée sans réserve, et il représente la majorité du corpus. Les niveaux 4 et 3 se distinguent par la force de l'engagement plutôt que par sa nature, « sans doute » pesant plus lourd que « il semble ». Au niveau 2, l'information devient une possibilité parmi d'autres, et c'est là que se rangent la plupart des « ? ». Le niveau 1 est réservé aux cas où le rédacteur signale lui-même que l'information est contestable.

L'exemple de la section précédente se relit dans ce cadre. Le « ? » y est une incertitude de niveau 2 dont la cible est le lieu du grade, le *vers* une imprécision dont la cible est la date, et l'obtention du grade reste au niveau 5, celui d'une information posée sans réserve.

Dans le graphe DAPHNÉ, la représentation visée est la suivante. Le factoïde de grade garde un attribut `certainty` sans réserve, son arête `TookPlaceAt` vers Paris reçoit un `certainty` qui marque le doute (niveau 2 de l'échelle) et son arête `OccurredAt` vers 1278 porte `datingtype = approximate`.

5.3 Évaluation sur les fiches Studium

L'évaluation porte sur 206 extraits de fiches, dont 158 expriment une incertitude et 48 n'en expriment aucune, pour 170 annotations de référence. Les 158 extraits incertains sont ceux des 191 éléments de la section précédente qui ont été conservés après relecture, 33 ayant été écartés, notamment des dates de forme purement technique et des doublons. Les 48 extraits sans incertitude ont été ajoutés pour mesurer les faux positifs. Un extrait peut en porter plusieurs, et les trois phénomènes distingués plus haut s'y cumulent librement. Sept modèles ont été comparés sur ce jeu. Chacun reçoit la même consigne, onze exemples résolus et l'extrait sous la forme de la figure 5.2, et doit rendre un objet JSON qui code chaque annotation par les cinq composantes de la section 5.2.

Les mesures comparent, extrait par extrait, l'ensemble prédit $\hat{A}(x)$ à la référence $A(x)$. Le tableau 5.2 rassemble les résultats. Le F1 est dit strict lorsque le phénomène et la catégorie sont justes tous les deux. La portée est évaluée par l'indice de Jaccard des ensembles de mots des deux segments, $J(s, \hat{s}) = |s \cap \hat{s}|/|s \cup \hat{s}|$, moyenné sur les paires. Le niveau est évalué par la part des paires dont le ℓ prédit égale celui de la référence. L'absence de niveau, qui concerne les 53 annotations d'imprécision et d'incomplétude, compte comme une valeur.

Modèle	Taille	Préc.	Rappel	F1 strict	Portée	Spécif.	Certitude
deepseek-v3.1	671 B	94,0	92,9	**93,5**	0,91	100	98,1
qwen3.5	397 B	94,4	88,8	91,5	0,90	100	98,0
Sonnet 4.6	—	88,7	92,4	90,5	0,91	97,9	98,1
gpt-oss-120b	120 B	91,0	89,4	90,2	0,91	100	97,4
qwen3-next	80 B	88,1	91,2	89,6	0,88	100	98,1
devstral	24 B	92,5	86,5	89,4	**0,92**	100	96,6
gpt-oss-20b	20 B	93,4	83,5	88,2	**0,92**	100	**98,6**

Table 5.2 – Résultats sur les 206 extraits, en pourcentages, sauf la portée, qui est un indice de Jaccard entre 0 et 1.

Six modèles sur sept ne produisent aucun faux positif sur les 48 extraits négatifs. Le septième en produit un. Ce faux positif unique éclaire une frontière du modèle d'annotation. La fiche d'ANDREAS Corsini porte, au champ des séjours, « Il a également séjourné auprès de son cousin à Avignon ». Le séjour est affirmé sans réserve et rien n'y est douteux. Sonnet 4.6 y voit pourtant une information manquante, au motif que l'identité de ce cousin n'est pas donnée. La remarque est exacte, ce cousin n'est effectivement pas nommé. Mais si le seul fait de ne pas tout préciser suffisait, chaque phrase du corpus serait incomplète, puisque « Il est né à Paris » ne dit pas à quelle adresse précise. Ce qui sépare les deux cas n'est pas la phrase, c'est le degré de précision attendu, et une base qui chercherait à reconstituer les réseaux familiaux compterait ce cousin anonyme comme une lacune. La décision de Sonnet 4.6 n'est pas fausse pour autant, elle répond à une exigence de précision que nous n'avions pas fixée.

La fiche G. Marie porte, au champ des grades, « Étudiant (Paris) 1466 ? ». L'annotation

de référence y voit une incertitude marquée par le « ? », de niveau 2, dont la portée est la proposition entière et la cible le statut d'étudiant. Quatre modèles sur sept, devstral, gpt-oss-120b, gpt-oss-20b et qwen3-next, rendent le même phénomène et le même niveau mais réduisent la portée à « 1466 » et la cible à la date. Les trois autres suivent la référence. Le désaccord tient à la position du signe. Le « ? » suit immédiatement l'année dans la valeur Studium, et ces quatre modèles le rattachent à ce qui le précède. Il y a une ambiguïté venant de la rédaction de la phrase, qu'il est impossible de résoudre sans information supplémentaire. Pour le graphe, les deux lectures n'affaiblissent pas la même arête. Avec la référence, le factoïde de grade entier porte le doute, et le fait même d'avoir été étudiant à Paris devient incertain. Avec la prédiction, seule l'arête `OccurredAt` vers 1466 le porte, le statut et le lieu restant certains. Le F1 strict compte cette prédiction comme fausse, alors que le phénomène et le niveau sont justes. Le même cas se présente sur quatre autres extraits du jeu.

La dernière colonne porte sur l'échelle de certitude du tableau 5.1. Elle inclut les annotations sans niveau, où une absence prédite compte comme juste, et aucun des sept modèles ne prédit de niveau là où la référence n'en a pas, ni n'en omet là où elle en a un. Sur les seules paires appariées dont la référence porte un niveau, entre 100 et 110 par modèle, le niveau exact est retrouvé dans 95,1 à 98,0 % des cas, et l'écart ne dépasse jamais un cran, sauf une fois pour gpt-oss-120b.

Comme au chapitre précédent, aucun de ces modèles ne se détache vraiment. Cinq points séparent le premier du dernier en F1 strict, tous trouvent le niveau de certitude exact plus de neuf fois sur dix. Désigner un vainqueur sur de tels écarts serait fragile, sur un jeu où quelques annotations font un point. La taille va bien dans le sens attendu, un modèle avec moins de paramètres faisant tendanciellement moins bien qu'un modèle plus grand.

5.4 L'accord entre modèles

Les scores agrégés ne disent pas si les sept modèles réussissent aux mêmes endroits. Le tableau 5.3 reprend donc la question binaire, l'extrait exprime-t-il une incertitude, et compte pour chacun des 206 extraits combien de modèles donnent la bonne réponse.

Modèles corrects	Extraits	Part
7 sur 7	188	91,3 %
6	5	2,4 %
5	5	2,4 %
4	2	1,0 %
3	3	1,5 %
2	1	0,5 %
1	1	0,5 %
0 sur 7	1	0,5 %

Table 5.3 – Nombre de modèles donnant la bonne réponse binaire, par extrait (206 extraits, sept modèles).

L'accord domine largement. Sur 188 des 206 extraits, les sept modèles répondent juste ensemble. Sur seize autres, ils se partagent entre la bonne et la mauvaise réponse. Les deux derniers extraits sont manqués, l'un par six modèles, l'autre par les sept, et le marqueur y est au même endroit. La fiche de B. Brianson porte « Paris » au champ de l'université, et c'est son commentaire qui écrit « Scribe qui semble avoir copié une lettre du recteur ». Celle de Wicardus Dierins donne pour médiane d'activité « 1265 », et son commentaire ajoute « Il est mort avant 1274 ». La référence annote le « semble » de l'un et le « avant » de l'autre. Dans les deux cas, le marqueur se trouve dans le commentaire du rédacteur et non dans la valeur du champ, ce qui constitue une erreur d'annotation, puisque les modèles devaient prendre en compte la valeur du champ uniquement. Les 51 erreurs restantes, soit 3,5 % des 1 442 réponses, se concentrent sur 18 extraits.

En production, la réponse de référence n'est pas disponible, seul le désaccord entre modèles est observable, et une erreur partagée à l'unanimité peut lui échapper. Ce signal permet donc de prioriser la relecture humaine, sans constituer une garantie.

5.5 Bilan du chapitre

Ce chapitre cherchait à détecter l'incertitude des notices et à la situer. Il en ressort un modèle à trois phénomènes, incertitude, imprécision et incomplétude, une échelle de certitude à cinq niveaux, et une portée qui désigne l'élément du graphe à affaiblir. Sur les 206 extraits du jeu de référence, les sept modèles de langage évalués obtiennent entre 88,2 et 93,5 % de F1 strict, retrouvent le niveau de certitude exact dans plus de 96 % des cas, et un seul faux positif a été observé sur les 48 extraits sans incertitude.

Le jeu de référence est toutefois petit, ce qui efforce à rendre prudent. Il est cependant déséquilibré vers des exemples incertains, ce qui joue plutôt en défaveur des modèles, puisqu'ils détectent presque parfaitement les extraits certains. Le faux positif observé rappelle en outre que le guide d'annotation laisse des questions ouvertes, comme le degré de précision attendu d'une phrase. Des ambiguïtés de rédaction, enfin, font que certains marqueurs

ne sont pas rattachés au bon élément du graphe. Sur ce jeu du moins, la détection paraît assez fiable pour dire si une phrase doute et où. Le chapitre 4 mesurait le degré seul, sur un autre jeu de phrases et avec un barème distinct. Ce chapitre y ajoute ce que le graphe demande, le phénomène et l'élément concerné.

Troisième partie

Bilan et perspectives

Chapitre 6

Discussion scientifique

les avis divergent entre les auteurs sur la chronologie des années suivantes ;

Studium Parisiense, fiche n° 22622, Petrus de Dacia (1)

6.1 Ce que les modèles ont montré

Un même constat revient d'une tâche à l'autre. Classer une institution, dire si une information est incertaine, imprécise ou incomplète, situer une phrase sur l'échelle de un à cinq, délimiter ce que le doute affecte, chaque fois quelques exemples donnés en consigne, dans le prompt, ont suffi pour que, sans aucun entraînement, les meilleurs modèles de langage s'approchent à quelques points du plafond de nos jeux d'incertitude.

Le niveau global de ces scores compte moins que l'endroit où ils se gagnent. La détection binaire de l'incertitude seule était presque acquise d'avance. Une liste de mots y approche des scores honorables. L'apport des modèles se loge dans ce que les listes ne savent pas faire, les phrases complexes, l'incertitude peu explicite, un « pouvoir » qui dit tantôt le doute et tantôt la capacité, et surtout la portée. Sur notre jeu de référence, où la réponse attendue n'est pas un segment de phrase mais l'élément précis que le doute affaiblit, les sept modèles évalués tiennent entre 88,2 et 93,5 % de F1 strict, métrique qui n'accorde le point que si le phénomène et la catégorie de l'élément visé sont tous deux corrects, et le niveau exact de certitude est retrouvé plus de 96 fois sur 100. Ils justifient de surcroît chaque décision en citant le marqueur qui la fonde, ce qui rend leurs erreurs relisibles.

Sans aucun entraînement toujours, ils égalent ou dépassent CamemBERT fine-tuné sur la quantification de l'incertitude. Nos essais montrent une sensibilité importante à la formulation du prompt. À modèle constant, reformuler la consigne et les exemples déplace les scores de plusieurs points, parfois de plusieurs dizaines. Une part du travail passe donc par écrire et tester ces prompts.

Deux remarques bornent ce tableau. Nous évaluons d'abord si le vote entre modèles améliore les résultats individuels. Sur le test du chapitre 4, le vote majoritaire des neuf modèles ne se trompe que sur 19 des 487 phrases, soit 96,1 % d'exactitude, au-dessus de sept des neuf votants mais à une phrase du meilleur, et sur la détection binaire du chapitre 5 il rend la bonne réponse pour 200 des 206 extraits. Des scores déjà proches de 100 laissent, il est vrai, peu à gagner, et l'on pourrait attendre que le vote rapporte davantage là où la marge est plus grande, la classification des institutions par exemple. Le tableau 3.1 répond que non, le vote des cinq modèles y obtient 86,4 % de F1, au-dessus de quatre votants mais sous les 87,6 % du meilleur. Le motif est donc le même partout. Le vote fait à peu près jeu égal avec le meilleur des modèles qu'il réunit, un peu au-dessous ici, un peu au-dessus là. Son intérêt pratique est ailleurs. Il atteint ce niveau sans qu'on ait à deviner d'avance quel modèle sera le meilleur sur son corpus, et ses rares désaccords avec le gold désignent les phrases à faire relire. La marge des institutions a d'ailleurs une explication propre. Décider si tel nom désigne un collège, un diocèse ou un ordre demande des connaissances fines, précises, d'histoire de France notamment, que les modèles n'ont pas tous rencontrées dans leurs données d'entraînement. Interrogés sans aucun accès à internet, sur leurs seuls paramètres, ils se trompent alors plus souvent. Ce savoir manquant peut alors se compléter à la main. La seconde remarque laisse de la marge. Aucun modèle frontière n'a été testé, les évaluations réunissent des modèles de taille moyenne à grosse, pour la plupart à poids ouverts, et ils saturent déjà presque nos jeux d'évaluation.

6.2 Ce que le jeu de référence décide

Ces scores ne se lisent pourtant qu'à travers le jeu de référence qui les produit, à commencer par sa règle de comptage. Une même sortie de détection peut en effet se noter de plusieurs façons. La plus lâche ne regarde que le verdict, la phrase est-elle spéculative, sans demander où le doute se loge. La deuxième compte mot à mot, chaque mot de la portée correctement délimité rapporte, un mot de trop ou de moins coûte, et une réponse presque juste garde presque tous ses points. La plus stricte raisonne en tout ou rien, le point n'est accordé que si le marqueur et la portée coïncident exactement avec le gold, et une réponse presque juste vaut zéro. Un exemple construit montre ce que chacune mesure. Sur « il a peut-être enseigné à Paris vers 1250 », la portée attendue est « enseigné à Paris » et la portée prédite « enseigné à Paris vers 1250 ». La règle du verdict seul tient la réponse pour juste, la phrase est bien spéculative. La règle mot à mot trouve les trois mots attendus et deux mots de trop, soit une précision de 3/5 pour un rappel de 1, un F1 de 0,75. La règle en tout ou rien compte zéro. La classification binaire et la détection de portée évaluent des sorties différentes, et ces trois règles ne sont pas trois degrés de sévérité interchangeables d'une même métrique. Le choix de la métrique est donc une décision importante pour l'évaluation, même si quelque peu futile en production.

Une part d'arbitraire ne se résorbe d'ailleurs jamais tout à fait. Choisir la granularité d'une échelle se discute, l'état de l'art en a rencontré de deux à neuf niveaux, voire à une échelle continue. Relier les mots aux nombres se discute, deux lecteurs d'un même « probable » n'entendent pas la même probabilité. Ranger une phrase dans un niveau se discute enfin, deux annotateurs consciencieux hésitent parfois entre 3 et 4 devant le même énoncé.

FactBank a rendu ce dernier geste presque mécanique par ses tests d'enchaînement, au prix d'une échelle grossière, certain, probable, possible. Affiner la granularité offre davantage d'expressivité au prix d'une part d'aléa, présente à chaque étape, et qui ne condamne pas l'exercice. Elle s'accepte. Un gold fixe un choix cohérent parmi d'autres défendables.

Avant de lire les scores d'un jeu de référence, il faut donc savoir d'où viennent ses étiquettes, de jugements agrégés à l'instinct ou de règles qu'un annotateur peut appliquer, si ceci alors cela. Le prompt donné aux modèles doit être précis, correct et complet (annexe A). Définir ainsi l'échelle de l'incertitude dans le prompt a fait passer le niveau exact de 33 à près de 98 % sur un benchmark. L'explicitation de la tâche a fortement amélioré les résultats dans cet essai.

Manquer la réponse attendue, enfin, n'est pas toujours manquer la bonne, car un jeu de référence ne fixe pas seulement ce qui est douteux, il fixe aussi la façon d'écrire la réponse. La portée en donne l'exemple le plus net. Tout le monde lit « il a peut-être enseigné à Paris » de la même façon, le doute porte sur l'enseignement à Paris. Mais quels mots faut-il surligner pour le noter ? « enseigné à Paris » seulement, ou aussi le « il » qui fait l'action ? Le sens ne change pas, seuls les mots retenus changent, et c'est le corpus qui décide, par convention, lesquels sont les bons. BioScope inclut parfois le sujet dans la portée, selon une règle qui dépend de la nature de ce sujet, et qu'un annotateur formé applique sans doute aisément. Il retient d'ailleurs partout l'empan syntaxique le plus large et y fait entrer le marqueur lui-même [76]. Les corpus cliniques français CAS et ESSAI tranchent en sens inverse, la portée y commence après le marqueur, qui en est exclu, « ces analogues pourraient [avoir un intérêt] » [19], si bien qu'une même réponse envoyée aux deux serait comptée juste ici et fausse là. Ces conventions locales, l'apprentissage supervisé les apprend dans ses données d'entraînement, mais l'avantage qu'il obtient là ne vaut pas nécessairement sur un autre benchmark. Écrire la convention dans le prompt envoyé à un modèle de langage tient en quelques lignes, et se réécrit pour le corpus suivant. C'est peut-être le principal avantage des modèles généralistes.

Pour ces raisons, les comparaisons entre corpus se lisent comme des indications, non comme un classement, tant que les protocoles ne sont pas alignés.

6.3 Le choix de l'échelle

Parmi ces décisions d'annotation, le choix de l'échelle est la décision la plus structurante, et le stage en a éprouvé deux. Le chapitre 4 note chaque phrase de 1 à 5 sur un axe unique, la certitude. Le chapitre 5 sépare trois axes de l'information, incertaine, imprécise ou incomplète, puis localise ce que le doute affecte. La seconde nous paraît mieux adaptée au graphe. « Il naît vers 1248 » est imprécis sans être douteux, et l'échelle unique force les nuances hétérogènes dans les mêmes niveaux. Trois axes séparés disent chacun ce qu'ils mesurent. L'échelle unique juge de plus la phrase entière, en lui attribuant un seul score même quand deux affirmations d'inégale solidité s'y côtoient, là où le schéma multi-axes note l'élément du graphe, et peut donc repérer plusieurs sources d'incertitude dans une même phrase.

6.4 De l’échantillon à la production

La principale limite d’exécution tient au périmètre. Il n’y a pas encore eu de passage en production. Tout ce qui précède a été mesuré sur des échantillons et des jeux de référence, quelques centaines de décisions à chaque fois, et le déploiement sur la base entière peut faire surgir des problèmes que ces essais n’ont pas rencontrés.

Ce passage à l’échelle devra aussi porter deux incertitudes de nature différente. Dans la terminologie de Jean (section 2.3), tout ce que ce rapport évalue au titre des erreurs de ses détecteurs et de ses classifieurs relève de l’incertitude *computationnelle*. Elle vient s’ajouter à l’incertitude *informationnelle* que les rédacteurs ont inscrite dans les notices. Le score de 1 à 5 encode le doute du rédacteur, les taux d’erreur de nos modèles encodent le nôtre, et un graphe complet doit pouvoir composer avec les deux.

Chapitre 7

Conclusions et perspectives

Paris : il arrive aux écoles vers 1120 ; on perd sa trace à partir de 1157.

Studium Parisiense, fiche n° 822, Adam de Parvo Ponte

7.1 Bilan du stage

Ce stage avait pour objectif de préserver l'incertitude exprimée dans les notices de Studium Parisiense lors de leur transformation en graphe de connaissances. Il s'agissait d'identifier les différentes formes qu'elle prend dans le corpus, de déterminer sur quels éléments elle porte, puis d'étudier comment la représenter et la quantifier. Le travail réalisé a donc porté à la fois sur la structuration des données, la caractérisation de l'incertitude et la mesure de son degré.

Un premier pipeline a permis de convertir un échantillon des fiches en un graphe suivant le modèle DAPHNÉ, dont l'élément central est le factoïde. La conversion repose sur des règles déterministes. En parallèle, des modèles de langage ont été évalués pour typer les institutions présentes dans les notices. Le meilleur modèle atteint un score F1 de 87,6 % sur les 139 entités du test, sans que ses prédictions aient été versées dans le graphe. Le graphe obtenu conserve pour chaque factoïde la notice et la rubrique dont il provient et fournit les emplacements nécessaires pour associer un niveau de certitude aux factoïdes ou aux relations concernées. Seul le « ? » accolé à une institutiony a laissé une première valeur, les scores de 1 à 5 obtenus aux chapitres 4 et 5 n'y ayant pas encore été versés.

Le premier volet sur le texte libre a porté sur la quantification de la certitude selon une échelle de 1 à 5. L'analyse du corpus utilisé dans l'article DKE a révélé des incohérences d'annotation, des doublons et des problèmes d'encodage. Les 2 234 phrases ont donc été réannotées. La réévaluation des méthodes montre que les meilleurs modèles de langage, utilisés sans entraînement spécifique sur ce corpus, obtiennent une exactitude comparable, voire supérieure, à celle de CamemBERT fine-tuné, ainsi que de meilleurs résultats

sur certains niveaux peu représentés. Ces résultats soulignent surtout l'importance de définir précisément la tâche et de disposer d'un jeu de référence cohérent avant de comparer les méthodes.

Le second volet a conduit à distinguer trois phénomènes : l'incertitude sur la validité d'un fait, l'imprécision de sa valeur et l'incomplétude de l'information. Cette distinction a été complétée par l'identification de la portée du doute, afin de déterminer s'il concerne l'énoncé entier, une relation ou une entité particulière. Sur le jeu de référence constitué à partir des fiches de Studium, les sept modèles évalués obtiennent, en F1 strict, des scores compris entre 88,2 % et 93,5 %, tandis que le niveau exact de certitude est retrouvé dans plus de 96 % des cas.

7.2 La suite

À court terme, trois chantiers peuvent démarrer. Le premier consiste à joindre les deux moitiés du stage, en versant les scores d'incertitude dans les attributs que les arêtes du graphe offrent déjà. Le deuxième passe en production la classification de la totalité des GroupP, des experts du domaine validant ou corrigeant les types que les modèles proposent. Le troisième complète le pipeline lui-même. Lire le volet `reference` de chaque bloc permettrait de rattacher chaque factoïde aux références bibliographiques de son bloc. Rapprocher les mentions d'une même personne au-delà de l'égalité stricte des noms, et étendre aux lieux les fusions faites sur les groupes, réduiraient les doublons. Les volets `value` et `comment` restent à exploiter, et les types Zone, GroupP et MoralEntity à harmoniser. Une relecture d'un échantillon de factoïdes contre les notices donnerait enfin une mesure de la qualité de la conversion, qui manque aujourd'hui.

Au-delà, plusieurs questions restent ouvertes. Le rapprochement des fiches décrivant un même individu sous des noms différents devra d'abord composer avec des affirmations qui portent chacune un degré. Deux notices ne se contredisent pas de la même manière selon qu'elles supposent ou qu'elles attestent, un cas que les cadres classiques du couplage d'enregistrements (section 2.9) ne couvrent pas. Les pistes envisagées combinent recherche contextuelle et réseaux de neurones de graphes.

La représentation de l'incertitude dans le graphe vient ensuite. Le stage a produit deux formes de réserve, un niveau de 1 à 5 sur une phrase, et une annotation qui distingue l'incertitude, l'imprécision et l'incomplétude et nomme l'élément visé. DAPHNÉ leur offre deux attributs, `certainty` et `datingtype`. Le projet de thèse qui prolonge ce stage sépare quatre dimensions, l'incomplétude, l'incertitude sur la vérité d'un fait, l'imprécision d'une valeur et la fiabilité de la source. Cette dernière n'est annotée nulle part dans ce travail, alors que chaque factoïde est rattaché à une source. Plusieurs formalismes se présentent pour porter ces dimensions séparément. Une date approximative, « vers 1248 », se décrit par un ensemble flou sur les années plutôt que par une borne nette. Les niveaux de 1 à 5 sont ordinaux, et la théorie des possibilités [20] les accueille sans la calibration qu'une probabilité exigerait. Les fonctions de croyance [69] combinent deux sources qui se contredisent et distinguent l'absence de preuve d'un fait tenu pour douteux. Reste à choisir où loger ces valeurs, en attributs d'arêtes comme aujourd'hui, en métadonnées de triplets avec RDF-

star [29], ou dans des plongements qui apprennent un score de confiance par triplet [16].

L'interrogation du graphe en dépend. Rien ne dit encore ce qu'il faut conclure d'un chemin dont deux arêtes portent un 3, et la réponse change selon l'opérateur retenu, que ce soit un minimum des degrés, un produit des probabilités ou une autre règle de combinaison des croyances. Le panorama de la gestion de l'incertitude dans la construction des graphes de connaissances balise ce terrain, de la fusion d'assertions contradictoires aux sémantiques d'agrégation [35]. Les méthodes qui projettent requêtes et entités dans un espace vectoriel tolèrent l'incomplétude du graphe et rendent chaque réponse avec un score, par des distributions bêta [61] ou par des opérations sur des ensembles flous [81]. Le raisonnement abductif sur graphe produit, à partir d'observations, les hypothèses logiques qui les expliquent [8], ce qui permet d'évaluer des hypothèses d'un historien, avec leur chemin de preuve. Aucune de ces approches n'a encore été évaluée sur des données prosopographiques. Les comparer suppose un jeu de référence de factoïdes annotés avec les historiens et des requêtes réelles, dont les 206 extraits du chapitre 5 sont une première forme. La fusion de plusieurs bases prosopographiques pose enfin ses propres difficultés. Le modèle DAPHNÉ a été conçu pour les accueillir toutes, et verser dans un même graphe Studium et des corpus voisins permettrait d'interroger d'un seul tenant des bases prosopographiques aujourd'hui séparées. Le rapprochement y change d'échelle, un même individu pouvant avoir sa fiche dans plusieurs bases, et chaque base arrive avec ses propres conventions du doute, qu'il faudrait réconcilier avant d'en comparer les degrés. Ces questions dessinent le programme de la thèse qui prolongera ce travail.

Annexe A

Le prompt donné aux modèles du chapitre 4

Voici le prompt donné aux neuf modèles de langage du chapitre 4, dont les résultats principaux apparaissent dans le tableau 4.3.

```
Tu es annotateur d'incertitude pour la base prosopographique Studium
(fiches biographiques d'universitaires médiévaux, en français). Tu
attribues à une phrase un score entier de 1 (fait tenu pour faux) à 5
(certain). Tu mesures l'engagement de l'auteur de la notice tel
qu'exprimé dans la langue, pas la vérité historique.

=== REGLE D'OR ===
Le score par défaut est 5. Tu ne descends sous 5 que si tu peux citer
mot pour mot un marqueur d'une des familles ci-dessous, lu en contexte.
Pas de marqueur citable = 5.

=== FAMILLE A - approximateurs de VALEUR -> 3 ===
« vers <date> », « aux alentours de », « environ », « à peu près »,
« presque », « plus ou moins » ; « avant <date> » / « après <date> »
(borne : la date exacte est inconnue) ; « entre <d1> et <d2> » QUAND
c'est l'estimation d'une date ponctuelle (naissance, mort, rédaction) ;
« <d1> ou <d2> » et toute alternative non résolue (« à X ou à Y »,
« oncle ou grand-oncle ») ; « plutôt » correctif ; « date estimée ».
EXCEPTIONS A :
- « près de / à côté de <lieu> » (localisation vague) -> 4 ; mais
  « Rouen, près de Paris » (apposition explicative) -> 5.
- DURÉE FACTUELLE -> 5 : un intervalle qui dénote une période réelle
  d'un état ou d'une charge (« prieure de Jouarre 1462-1479 »,
  « évêque 1351-1361 », « au service du roi 1350-1368 », « étudiant
  1266-1277 »). Test : événement ponctuel par nature (naissance, mort,
  élection) -> l'intervalle est une estimation -> 3 ; état duratif
  (charge, épiscopat, résidence) -> durée réelle -> 5 ; processus
  ambigu (rédaction, dictée) -> 3.

=== FAMILLE B - plausibilité / possibilité ===
« semble », « paraît », « probablement », « sans doute »,
« vraisemblablement », « apparemment », « jugé probable » -> 4
(« sans doute » = probablement -> 4, JAMAIS 5).
« peut-être », « possiblement », « il n'est pas
impossible que », « hypothèse » -> 3.
Conditionnel ÉPISTÉMIQUE (« il serait né », « aurait été ») -> 3 ;
```

```
mais futur du passé ou irréel (« il annonça qu'il serait présent »)
-> PAS un marqueur -> 5.
« si l'on identifie », « s'il s'agit » -> 3.
« pouvoir » capacitaire (« il peut lire le grec ») -> PAS un
marqueur -> 5.

=== FAMILLE C - savoir rapporté -> 4 ===
« selon X », « d'après X », « pour X, ... », « il passe pour »,
« on dit que », « on prétend », « il est décrit comme »,
« cité par X », « attesté par X », « mentionné par X », « signalé par
X ». Jamais 5 dès qu'un rapporteur est nommé.
« on a AUSSI suggéré / il est AUSSI dit » (hypothèse concurrente) -> 3.
MAIS « cité DANS <une œuvre/un manuscrit> » = localisation
bibliographique -> 5.

=== FAMILLE D - attribution d'œuvre ===
« attribué(e) à X » neutre, traditionnel, ou « longtemps attribuée »
sans rejet -> 4.
« aussi / parfois attribué(e) à », « attribué à X ou à Y », « attribué
par certains » -> 3.
« attribution douteuse / non assurée / contestée / jugée incertaine »,
ou attribution + contre-évidence explicite -> 2.
« faussement attribué », « pseudépigraphe », « attribution rejetée /
très douteuse » -> 1.
MAIS assertion directe sans mot d'attribution (« il a été écrit par
X ») -> 5.

=== FAMILLE E - interrogatives ===
« ? » seul (y compris « ? » en tête de fiche, « 1352 ? ») -> 3.

=== FAMILLE F - doute déclaré ===
« incertain(e) » simple -> 3.
« douteux/douteuse » simple, « n'est pas assuré », « n'est pas
certain », « pas sûr », « mis en doute », « sujette à caution » -> 2.
« très douteux », « plus que douteux », « incertitude totale »,
« authenticité rejetée », « apocryphe », « si elle a jamais
existé » -> 1.

=== FAMILLE G - ignorance déclarée -> 2 ===
« datation arbitraire », « date conjecturale/conjecturelle »,
« on ignore », « on ne sait pas », « date inconnue ».

=== CE QUI N'EST PAS UN MARQUEUR (score 5) ===
- Nom manquant « X », « N. », « N.... » : la relation est certaine,
  seul le libellé manque -> 5 (sauf marqueur explicite : « sa femme
  est peut-être X » -> 3).
- « Œuvre incomplète / perdue / ne subsiste que des fragments » =
  constat factuel -> 5 (sauf marqueur : « sans doute perdue » -> 4).
- Négation qui CORRIGE vers un fait ferme (« il ne s'agit pas de A ...
  qui sont de B ») -> 5 ; mais fait consigné REJETÉ (« le commentaire
  dit du Pseudo-X n'est pas de X ») -> 1.
- Lieux, références bibliographiques neutres, nombres, dates précises,
  liens de parenté attestés.

=== REGLE DE CUMUL ===
Score = le MINIMUM des scores des marqueurs présents. Une seule exception : un
indicateur + « ? » portant sur le même
fait -> 2 (« Son cousin est peut-être Jean M. ... ? » -> 2).

=== PROTOCOLE DE RÉPONSE ===
1. Cite entre guillemets le ou les marqueurs présents avec leur
   famille (ou « aucun marqueur citable »). Une à deux lignes maximum.
2. Applique le minimum (+ exception éventuelle).
3. Termine strictement par l'objet JSON, et rien après.
```

```
=== EXEMPLES ===
Phrase : « Son père est Jean de SCEPPERE, tailleur. »
Marqueurs : aucun marqueur citable. {"score": 5}

Phrase : « Sa femme est X, fille de Guillaume BASIN. »
Marqueurs : aucun (« X » = nom manquant, la relation est certaine).
{"score": 5}

Phrase : « Elle est abbesse de Jouarre en 1330-1342. »
Marqueurs : aucun (« 1330-1342 » = durée factuelle d'une charge).
{"score": 5}

Phrase : « Il est mort vers 1187. »
Marqueurs : « vers 1187 » (A). {"score": 3}

Phrase : « L'œuvre est attribuée à Petrus de ANGLIA. »
Marqueurs : « attribuée à » (D neutre). {"score": 4}

Phrase : « Œuvre aussi attribuée à ALBERT le Grand. »
Marqueurs : « aussi attribuée à » (D concurrente). {"score": 3}

Phrase : « Selon THIOLIER, il est l'auteur du traité. »
Marqueurs : « Selon THIOLIER » (C). {"score": 4}

Phrase : « L'authenticité de cette œuvre n'est pas assurée. »
Marqueurs : « n'est pas assurée » (F). {"score": 2}

Phrase : « Œuvre douteuse ou pseudépigraphe. »
Marqueurs : « pseudépigraphe » (D rejetée / F niveau 1). {"score": 1}

Phrase : « Son cousin est peut-être Jean MARTIN, chanoine de Reims ? »
Marqueurs : « peut-être » (B, 3) + « ? » (E) sur le même fait ->
exception -> 2. {"score": 2}

Pour la phrase qui suit : cite le ou les marqueurs avec leur famille
(ou « aucun marqueur citable »), applique la règle du minimum, puis
termine par {"score": <1 à 5>}.
```

Bibliographie

[1] Ran Abramitzky et al. « Automated Linking of Historical Data ». In : *Journal of Economic Literature* 59.3 (2021), p. 865-918. DOI : 10.1257/jel.20201599.

[2] Jacky Akoka, Isabelle Comyn-Wattiau et Cédric du Mouza. « Conception de bases de données prosopographiques en histoire – un état de l'art ». In : *Revue ouverte d'ingénierie des systèmes d'information* 1.3 (2020), p. 1-19. DOI : 10.21494/ISTE.OP.2020.0531. HAL : hal-03937727.

[3] Jacky Akoka et al. « Conceptual Modeling of Prosopographic Databases Integrating Quality Dimensions ». In : *Journal of Data Mining & Digital Humanities* (2021) : *Special Issue on Data Science and Digital Humanities @EGC 2018*. DOI : 10.46298/jdmdh.5078. HAL : hal-01966374.

[4] Jacky Akoka et al. « Contribution of Conceptual Modeling to Enhancing Historians' Intuition – Application to Prosopography ». In : *Conceptual Modeling – 39th International Conference (ER 2020)*. Sous la dir. de Gillian Dobbie et al. T. 12400. Lecture Notes in Computer Science. Springer, 2020, p. 164-173. DOI : 10.1007/978-3-030-62522-1_12. arXiv : 2011.13276.

[5] Emilia Apostolova, Noriko Tomuro et Dina Demner-Fushman. « Automatic Extraction of Lexico-Syntactic Patterns for Detection of Negation and Speculation Scopes ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011, p. 283-287. URL : https://aclanthology.org/P11-2049/.

[6] Iana Atanassova, François-C. Rey et Marc Bertin. « Studying Uncertainty in Science: A Distributional Analysis through the IMRaD Structure ». In : *WOSP: 7th International Workshop on Mining Scientific Publications at the 11th Edition of the Language Resources and Evaluation Conference*. 2018.

[7] Naser Ayat et al. « Entity Resolution for Probabilistic Data ». In : *Information Sciences* 277 (2014), p. 492-511. DOI : 10.1016/j.ins.2014.02.135.

[8] Jiaxin Bai et al. « Advancing Abductive Reasoning in Knowledge Graphs through Complex Logical Hypothesis Generation ». In : *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. 2024, p. 1312-1329.

[9] Gerrit Bloothooft et al., éd. *Population Reconstruction*. Cham : Springer, 2015. DOI : 10.1007/978-3-319-19884-2.

[10] John Bradley. « A Prosopography as Linked Open Data: Some Implications from DPRR ». In : *Digital Humanities Quarterly* 14.2 (2020). DOI : 10.63744/2aff26m7ec7d.

[11] John Bradley et Harold Short. « Texts into Databases: The Evolving Field of New-Style Prosopography ». In : *Literary and Linguistic Computing* 20.Suppl. 1 (2005), p. 3-24. DOI : 10.1093/llc/fqi022.

[12] Tom B. Brown et al. « Language Models Are Few-Shot Learners ». In : *Advances in Neural Information Processing Systems*. T. 33. 2020, p. 1877-1901.

[13] Wendy W. Chapman, David Chu et John N. Dowling. « ConText: An Algorithm for Identifying Contextual Features from Clinical Text ». In : *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics, 2007, p. 81-88. URL : https://aclanthology.org/W07-1011/.

[14] Chao Chen, Min Song et Go Eun Heo. « A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty ». In : *Journal of Informetrics* 12.1 (2018), p. 158-180. DOI : 10.1016/j.joi.2017.12.004.

[15] Song Chen et Hongsu Wang. « China Biographical Database (CBDB): A Relational Database for Prosopographical Research of Pre-Modern China ». In : *Journal of Open Humanities Data* 8.4 (2022), p. 1-6. DOI : 10.5334/johd.68.

[16] Xuelu Chen et al. « Embedding Uncertain Knowledge Graphs ». In : *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*. 2019, p. 3363-3370.

[17] Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg : Springer, 2012. DOI : 10.1007/978-3-642-31164-2.

[18] Paulo C. G. Costa et al. « Towards Unbiased Evaluation of Uncertainty Reasoning: The URREF Ontology ». In : *Proceedings of the 15th International Conference on Information Fusion (FUSION 2012)*. IEEE, 2012, p. 2301-2308. URL : https://ieeexplore.ieee.org/document/6290584/.

[19] Clément Dalloux, Vincent Claveau et Natalia Grabar. « Speculation and Negation Detection in French Biomedical Corpora ». In : *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. 2019, p. 223-232. DOI : 10.26615/978-954-452-056-4_026.

[20] Didier Dubois et Henri Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. New York : Plenum Press, 1988.

[21] Richárd Farkas et al. « The CoNLL-2010 Shared Task: Learning to Detect Hedges and Their Scope in Natural Language Text ». In : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*. Association for Computational Linguistics, 2010, p. 1-12. URL : https://aclanthology.org/W10-3001/.

[22] Ivan P. Fellegi et Alan B. Sunter. « A Theory for Record Linkage ». In : *Journal of the American Statistical Association* 64.328 (1969), p. 1183-1210. DOI : 10.1080/01621459.1969.10501049.

[23] Bruno de Finetti. « La prévision : ses lois logiques, ses sources subjectives ». In : *Annales de l'institut Henri Poincaré* 7.1 (1937), p. 1-68. URL : https://www.numdam.org/item/AIHP_1937__7_1_1_0/.

[24] Jean-Philippe Genet. « Studium Parisiense, un répertoire informatisé des écoles et de l'université de Paris ». In : *Annali di Storia delle università italiane* XXI.1 (2017), p. 25-74. HAL : halshs-02972204.

[25] Jean-Philippe Genet, Thierry Kouamé et Stéphane Lamassé. « The Socio-Economic Role of Medieval Parisian Colleges through the "Studium Parisiense" Database ». In : *CIAN-Revista de Historia de las Universidades* 24.1 (2021), p. 82-125. doi : 10.20318/cian.2021.6159.

[26] Jean-Philippe Genet et al. « General Introduction to the Studium Project ». In : *Medieval Prosopography* 31 (2016), p. 155-170. HAL : halshs-02987904.

[27] Kleanthi Georgala et al. « Record Linkage in Medieval and Early Modern Text ». In : *Population Reconstruction*. Sous la dir. de Gerrit Bloothooft et al. Cham : Springer, 2015, p. 173-195. doi : 10.1007/978-3-319-19884-2_9.

[28] Nicolas Gutehrlé, Panggih Kusuma Ningrum et Iana Atanassova. « A Large-Scale Multi-Disciplinary Analysis of Uncertainty in Research Articles ». In : *Data & Knowledge Engineering* 163, 102561 (2026). doi : 10.1016/j.datak.2026.102561.

[29] Olaf Hartig et Bryan Thompson. *Foundations of an Alternative Approach to Reification in RDF*. 2014. arXiv : 1406.3399.

[30] Aidan Hogan et al. « Knowledge Graphs ». In : *ACM Computing Surveys* 54.4 (2021), 71:1-71:37. doi : 10.1145/3447772.

[31] Laurence R. Horn. *A Natural History of Negation*. Chicago : University of Chicago Press, 1989.

[32] Leo Hoye. *Adverbs and Modality in English*. Longman English Language Series. London : Longman, 1997.

[33] Ken Hyland. *Hedging in Scientific Research Articles*. Pragmatics & Beyond New Series 54. Amsterdam : John Benjamins, 1998.

[34] Eero Hyvönen. « Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery ». In : *Semantic Web* 11.1 (2020), p. 187-193.

[35] Lucas Jarnac, Yoan Chabot et Miguel Couceiro. « Uncertainty Management in the Construction of Knowledge Graphs: A Survey ». In : *Transactions on Graph Data and Knowledge* 3.1 (2025), 3:1-3:48. doi : 10.4230/TGDK.3.1.3.

[36] Pierre-Antoine Jean. « Gestion de l'incertitude et de l'imprécision dans un processus d'extraction de connaissances à partir des textes ». Thèse de doct. Université de Montpellier, 2017. HAL : tel-01905221. NNT : 2017MONTS019.

[37] Sherman Kent. « Words of Estimative Probability ». In : *Studies in Intelligence* 8.4 (1964).

[38] Fadhela Kerdjoudj et Olivier Curé. « Evaluating Uncertainty in Textual Document ». In : *Proceedings of the 11th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2015), co-located with ISWC 2015*. T. 1479. CEUR Workshop Proceedings. 2015, p. 1-13. URL : https://ceur-ws.org/Vol-1479/paper1.pdf.

[39] Aditya Khandelwal et Benita Kathleen Britto. « Multitask Learning of Negation and Speculation Using Transformers ». In : *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis (LOUHI 2020)*. Association for Computational Linguistics, 2020, p. 79-87. DOI : 10.18653/v1/2020.louhi-1.9.

[40] Wissam Mammar Kouadri et al. « Uncertainty Detection in Historical Databases ». In : *Natural Language Processing and Information Systems (NLDB 2022)*. T. 13286. Lecture Notes in Computer Science. Springer, 2022, p. 73-85. DOI : 10.1007/978-3-031-08473-7_7. HAL : hal-03843641.

[41] Thierry Kouamé et Stéphane Lamassé. « The Quest for Origins. Academic Filiations in the Studium Parisiense Database ». In : *Specimina Nova* 13 (2024), p. 47-61. HAL : hal-05436611.

[42] Florian Kräutli et Stephen Boyd Davis. « Known Unknowns: Representing Uncertainty in Historical Time ». In : *Proceedings of Electronic Visualisation and the Arts (EVA London 2013)*. BCS, 2013, p. 61-68.

[43] Xue Li. « From Fine-Tuning to Prompting: A Paradigm Shift in Knowledge Graph Construction ». Thèse de doct. Universiteit van Amsterdam, 2026. SIKS Dissertation Series No. 2026-12.

[44] Stephanie Lin, Jacob Hilton et Owain Evans. « Teaching Models to Express Their Uncertainty in Words ». In : *Transactions on Machine Learning Research* (2022). arXiv : 2205.14334.

[45] Yao Lu et al. « Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity ». In : *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, p. 8086-8098. DOI : 10.18653/v1/2022.acl-long.556.

[46] Ahmed Mahany et al. « Negation and Speculation in NLP: A Survey, Corpora, Methods, and Applications ». In : *Applied Sciences* 12.10, 5209 (2022). DOI : 10.3390/app12105209.

[47] Marie-Catherine de Marneffe, Christopher D. Manning et Christopher Potts. « Did It Happen? The Pragmatic Complexity of Veridicality Assessment ». In : *Computational Linguistics* 38.2 (2012), p. 301-333. DOI : 10.1162/COLI_a_00097.

[48] Louis Martin et al. « CamemBERT: A Tasty French Language Model ». In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, p. 7203-7219. DOI : 10.18653/v1/2020.acl-main.645. HAL : hal-02889805.

[49] Patricia Martín-Rodilla et Cesar Gonzalez-Perez. « Representing Imprecise and Uncertain Knowledge in Digital Humanities: A Theoretical Framework and ConML Implementation with a Real Case Study ». In : *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2018)*. ACM, 2018, p. 863-871. DOI : 10.1145/3284179.3284318.

[50] Patricia Martín-Rodilla, Martín Pereira-Fariña et Cesar Gonzalez-Perez. « Qualifying and Quantifying Uncertainty in Digital Humanities: A Fuzzy-Logic Approach ». In : *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2019)*. ACM, 2019, p. 788-794. DOI : 10.1145/3362789.3362833.

[51] Kamil Matoušek, Martin Falc et Zdeněk Kouba. « Extending Temporal Ontology with Uncertain Historical Time ». In : *Computing and Informatics* 26.3 (2007), p. 239-254.

[52] Ben Medlock et Ted Briscoe. « Weakly Supervised Learning for Hedge Classification in Scientific Literature ». In : *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*. Association for Computational Linguistics, 2007, p. 992-999. URL : https://aclanthology.org/P07-1125/.

[53] Howard B. Newcombe et al. « Automatic Linkage of Vital Records ». In : *Science* 130.3381 (1959), p. 954-959. DOI : 10.1126/science.130.3381.954.

[54] Panggih Kusuma Ningrum, Philipp Mayr et Iana Atanassova. « UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text ». In : *Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE) and the 3rd AI + Informetrics (AII), EEKE-AII 2023*. 2023. arXiv : 2307.14236.

[55] Panggih Kusuma Ningrum et al. « Annotating Scientific Uncertainty: A Comprehensive Model Using Linguistic Patterns and Comparison with Existing Approaches ». In : *Journal of Informetrics* 19.2, 101661 (2025). DOI : 10.1016/j.joi.2025.101661. arXiv : 2503.11376.

[56] Michele Pasin et John Bradley. « Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach ». In : *Digital Scholarship in the Humanities* 30.1 (2015), p. 86-97. DOI : 10.1093/llc/fqt037.

[57] Zdzisław Pawlak. « Rough Sets ». In : *International Journal of Computer and Information Sciences* 11.5 (1982), p. 341-356. DOI : 10.1007/BF01001956.

[58] Quentin Pradet, Gaël de Chalendar et Jeanne Baguenier Desormeaux. « WoNeF, an Improved, Expanded and Evaluated Automatic French Translation of WordNet ». In : *Proceedings of the Seventh Global Wordnet Conference (GWC)*. 2014, p. 32-39.

[59] Ellen F. Prince, Joel Frader et Charles Bosk. « On Hedging in Physician-Physician Discourse ». In : *Linguistics and the Professions*. Sous la dir. de Robert J. Di Pietro. Norwood, NJ : Ablex, 1982, p. 83-97.

[60] Giulio Quaresima et Stefania Zucchini. « "Onomasticon, Prosopografia dell'Università degli Studi di Perugia": The Origin of the Project, Present Database Features and Future Challenges ». In : *Specimina Nova Pars Prima Sectio Mediaevalis* 13 (2024), p. 117-128. doi : 10.15170/SPMNNV.2024.13.05.

[61] Hongyu Ren et Jure Leskovec. « Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs ». In : *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. 2020, p. 19716-19726.

[62] François-Claude Rey. « Ontologie de l'incertitude et annotation sémantique d'un corpus autour du changement climatique ». Thèse de doct. Université Bourgogne Franche-Comté, 2022. HAL : tel-03780719. NNT : 2022UBFCC007.

[63] Victoria L. Rubin. « Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements ». In : *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. 2007, p. 141-144.

[64] Victoria L. Rubin, Elizabeth D. Liddy et Noriko Kando. « Certainty Identification in Texts: Categorization Model and Manual Tagging Results ». In : *Computing Attitude and Affect in Text: Theory and Applications*. Springer, 2006, p. 61-76.

[65] Rachel Rudinger, Aaron Steven White et Benjamin Van Durme. « Neural Models of Factuality ». In : *Proceedings of NAACL-HLT 2018, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, p. 731-744. doi : 10.18653/v1/N18-1067.

[66] Françoise Salager-Meyer. « Hedges and Textual Communicative Function in Medical English Written Discourse ». In : *English for Specific Purposes* 13.2 (1994), p. 149-170.

[67] Roser Saurí et James Pustejovsky. « FactBank: A Corpus Annotated with Event Factuality ». In : *Language Resources and Evaluation* 43.3 (2009), p. 227-268. doi : 10.1007/s10579-009-9089-9.

[68] Daniel L. Schwartz, Nathan P. Gibson et Katayoun Torabi. « Modeling a Born-Digital Factoid Prosopography Using the TEI and Linked Data ». In : *Journal of the Text Encoding Initiative* (Rolling Issue 2022). doi : 10.4000/jtei.3979.

[69] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton : Princeton University Press, 1976.

[70] Andrzej Skowron et Jarosław Stepaniuk. « Tolerance Approximation Spaces ». In : *Fundamenta Informaticae* 27.2-3 (1996), p. 245-253. doi : 10.3233/FI-1996-272311.

[71] Theodore C. Sorensen. *Kennedy*. New York : Harper & Row, 1965.

[72] TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.12.0. Text Encoding Initiative Consortium. 2026. URL : https://tei-c.org/release/doc/tei-p5-doc/en/html/CE.html.

[73] Katherine Tian et al. « Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback ». In : *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, p. 5433-5442.

[74] Amos Tversky. « Features of Similarity ». In : *Psychological Review* 84.4 (1977), p. 327-352.

[75] Veronika Vincze. « Uncertainty Detection in Natural Language Texts ». Thèse de doct. Szeged, Hongrie : Szegedi Tudományegyetem, 2015. DOI : 10.14232/phd.2291.

[76] Veronika Vincze et al. « The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and Their Scopes ». In : *BMC Bioinformatics* 9.Suppl 11, S9 (2008). DOI : 10.1186/1471-2105-9-S11-S9.

[77] Thomas S. Wallsten et al. « Measuring the Vague Meanings of Probability Terms ». In : *Journal of Experimental Psychology: General* 115.4 (1986), p. 348-365. DOI : 10.1037/0096-3445.115.4.348.

[78] Edward A. Wrigley et Roger S. Schofield. « Nominal Record Linkage by Computer and the Logic of Family Reconstitution ». In : *Identifying People in the Past*. Sous la dir. d'Edward A. Wrigley. London : Edward Arnold, 1973, p. 64-101.

[79] Wei Zhang et al. « Visual Reasoning for Uncertainty in Spatio-Temporal Events of Historical Figures ». In : *IEEE Transactions on Visualization and Computer Graphics* 29.6 (2023), p. 3009-3023. DOI : 10.1109/TVCG.2022.3146508.

[80] Zihao Zhao et al. « Calibrate Before Use: Improving Few-Shot Performance of Language Models ». In : *Proceedings of the 38th International Conference on Machine Learning*. T. 139. Proceedings of Machine Learning Research. PMLR, 2021, p. 12697-12706.

[81] Zhaocheng Zhu et al. « Neural-Symbolic Models for Logical Queries on Knowledge Graphs ». In : *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*. 2022, p. 27454-27478.