Université Paris III Sorbonne Nouvelle

Assimilation de triplets obtenus par extracteurs d'information libre à des faits de bases de connaissances

par
David ALFONSO HERMELO

Département Institut de Linguistique et Phonétique Générales et Appliquées, ILPGA
UFR Littérature, linguistique, didactique (LLD)

Mémoire présenté à la UFR Littérature, linguistique, didactique (LLD)
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en Traitement Automatique des Langues

Septembre, 2017

Université Paris III Sorbonne Nouvelle
UFR Littérature, linguistique, didactique (LLD)

Ce mémoire intitulé:

Assimilation de triplets obtenus par extracteurs d'information libre à des faits de bases de connaissances

présenté par:

David ALFONSO HERMELO

a été évalué par un jury composé des personnes suivantes:

Isabelle TELLIER ,	président-rapporteur
Philippe LANGLAIS ,	directeur de recherche
Sylvain KAHANE ,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

Les bases de connaissances sont des systèmes de stockage d'informations factuelles (des faits) sur le monde. Elles permettent à des systèmes automatiques d'interagir avec les utilisateurs humains, de faciliter le traitement de l'information, de faire des associations de concepts.

Les compagnies qui développent les bases de connaissances s'intéressent pas ou peu à stocker autre chose que des faits objectifs et vérifiables.

Les extracteurs d'information libre permettent de transformer le contenu d'un document écrit en langage naturel en triplets minimaux (sujet-propriété-objet).

Dans ce mémoire nous faisons un premier pas vers l'analyse comparative entre les faits des bases de connaissances et les triplets obtenus avec des extracteurs d'information libre. Nous faisons également des expériences sur la manière d'assimiler les triplets extraits en vue d'une possible comparaison entre triplets et faits de bases de connaissances.

Mots clés: Traitement automatique des langues, extraction d'information, base de connaissances, extracteurs d'information libre, triplets, faits.

ABSTRACT

Knowledge bases are systems for storing factual information (facts) about the world. They allow automatic systems to interact with human users, to facilitate information processing, to make conceptual associations.

Companies that develop knowledge bases have little or no interest in storing anything other than objective and verifiable facts.

The open information extractors allow to transform the content of a document written in natural language into minimal triplets (subject-property-object).

In this paper we take a first step towards the comparative analysis between the facts of the knowledge bases and the triplets obtained with the open information extractors. We also do experiments on how to assimilate the extracted triplets for a possible comparison between triplets and knowledge base facts.

Keywords: Natural language processing, information extraction, knowledge base, open information extraction, triples, facts, OIE, KB.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES SIGLES

NOTA : Certains de ces sigles sont en langue anglaise par convention.

API	Application program interface (Interface de programme d'application)
AUC	Area Under the Curve (Aire Sous la Courbe)
E	Entité d'entrée d'une requête de KB ou étant le thème de l'information libre fournie en entrée d'une OIE
E'	Coréférence de E
IA	Intelligence Artificielle
KB	Knowledge Base (Base de Connaissances)
KG	Knowledge Graph (Base de Connaissances de Google)
O	Objet du triplet/fait, argument 2 du triplet/fait
OIE	Open Information Extractor (Extracteurs d'Information Libre)
P	Propriété du triplet/fait, prédicat du triplet/fait, relation du triplet/fait
RDF	Resource Description Framework
S	Sujet du triplet/fait, argument 1 du triplet/fait
S-P-O	Sujet-Propriété-Objet
SRL	Semantic Role Labeling

(dédicace) A Oscar, Lourdès et Sergio.

REMERCIEMENTS

À Philippe Langlais pour son encadrement, son soutien, son aide et sa patience. Aux professeurs de la formation PluriTAL qui m'ont fait aimer cette spécialité, nuits blanches inclues. À mes camarades et amis de la formation pour les bons moments passés ensemble et les mauvais moments endurés en groupe. À ma famille pour leur aide et soutien constant. À mes camarades du laboratoire RALI de l'Université de Montréal pour leur aide précieuse et leurs conseils. À Gabriel Stanovsky pour ses réponses rapides à nos questions.

CHAPITRE 1

INTRODUCTION

1.1 Observation

Les bases de connaissances (KB[1]) "encyclopédiques" regroupent des faits (informations factuelles immuables, objectives, neutres, populairement acceptées) d'origines et de types très divers sur le plus grand nombre d'entités (concepts sémantiques) possible. Traditionnellement, ces faits sont entreposées sous forme de triplets (S-P-O[2]) souvent associés à une ontologie.

Elles sont largement utilisées en Intelligence Artificielle (IA), dans des systèmes de web sémantique et dans des systèmes d'assistant numérique.

Elles permettent la vérification informative, le résumé factuel, l'annotation de textes bruts et peut fournir l'introduction d'un semblant de valeur sémantique au texte en analysant les interconnections des entités.

Il existe une demande grandissante de KB incluant de plus en plus d'entités et de faits. Mais tandis que de nombreux groupes de recherche et entreprises font la course à l'information factuelle, les informations subjectives sont négligées.

Une méthode courante pour peupler les KB (ajouter des faits) est d'utiliser des systèmes d'extraction d'information libre (OIE [3]) pour transformer des documents structurés ou semi-structurés en triplets, puis d'en faire une vérification ou édition humaine avant de les inclure définitivement.

[1] "Knowledge Base".
[2] sujet-propriété-objet
[3] "Open Information Extractors"

1.2 Problématique de recherche

Bien que les faits des KB et les triplets de sortie OIE [4] aient une structure similaire, il existe peu de comparaisons et analyses statistiques entre les deux. Nous ne nions pas qu'il y a de nombreux obstacles qui rendent leur comparaison et analyse difficiles :

- leur entrée ("input") n'est pas de même nature,
- les faits de sortie ("output") des KB sont pertinents à la requête alors que la pertinence des triplets de sortie OIE est instable (selon l'OIE et selon chaque phrase de chaque document d'entrée, le sujet du triplet (S) peut ne pas correspondre à la requête),
- les faits de sortie des KB sont fiables (dans le sens de véridiques, authentiques, sémantiquement valides, fidèles à la réalité) alors que la fiabilité des triplets de sortie OIE dépend de chaque document d'entrée,
- les faits de sortie des KB sont pour la plus grande part objectifs [5] alors que les triplets de sortie OIE sont pour la plus grande part subjectifs [6],
- le sujet des faits de sortie des KB est uniforme alors que les sujets des triplets de sortie OIE sont nombreux et variés.

Devant ces obstacles, nous nous sommes demandés s'il existait un moyen de les éliminer. Du moins suffisamment pour que les faits de KB et les triplets de sortie OIE

[4] Nous utilisons le terme "triplet de sortie OIE" pour nous référer à l'ensemble de triplets extraits à partir d'information libre moyennant un ou plusieurs OIE (c'est-à-dire aux triplets obtenus comme sortie ("output") d'un outil OIE lorsque nous lui fournissons comme entrée ("input") de l'information libre). Nous utilisons ce terme afin d'avoir un moyen simple et concret de distinguer entre ce type spécifique de triplets et des faits de KB, et les triplets ontologiques, et le concept même de triplet, etcætera.

[5] Sans trop nous aventurer dans la métaphysique, nous entendons par "élément objectif" une information vérifiable, constante et acceptée par un public majoritaire.

[6] Nous entendons par "élément subjectif" une information non vérifiable, éphémère, incertaine et non vérifiée (ou pas encore vérifiée). Cela implique qu'une information subjective peut correspondre à une information objective mais tant que celle-ci n'aura pas été le sujet d'une vérification elle ne peut pas être classifiée comme "objective".

soient comparables. Comme les faits de KB sont bien plus invariables que les triplets de sortie OIE, nous avons décidé de concentrer nos efforts à **modifier les triplets de sortie OIE jusqu'à les assimiler le plus possible à des faits de KB**.

Pour ce faire, nous nous sommes proposés plusieurs étapes nécessaires pour aboutir à notre objectif final :

- faire une sélection des KB les plus complètes,
- analyser en profondeur les KB sélectionnées,
- faire une sélection des OIE les plus compétents et les mieux adaptés à notre projet,
- analyser en profondeur les sorties des OIE sélectionnés,
- analyser l'état de l'art et les projets similaires,
- faire des expériences comparatives entre des triplets de sortie OIE et des faits de KB,
- conceptualiser des heuristiques simples dans le but de capturer un maximum de triplets de sortie OIE similaires aux faits de KB,
- faire des expériences quantitatives et qualitatives sur les triplets de sortie OIE capturés avec les heuristiques,
- analyser, commenter et conclure sur les résultats des expériences.

Cette tâche d'assimilation n'est pas un problème trivial. C'est pourquoi nous précisons que nous allons tenter d'assimiler au mieux les triplets de sortie OIE et les faits de KB mais qu'il faudra certainement perfectionner les méthodes, en trouver des nouvelles et faire davantage d'évaluations dans des travaux ultérieurs.

1.3 Utilité éventuelle et prospective

Les KB actuelles contiennent très peu de faits subjectifs (voire aucun).

Cette limitation des KB aux faits objectifs est justifiée par le fait qu'il est difficile de juger de la validité d'une information subjective sans risquer un jugement biaisé (ce qui est vrai pour un vérificateur humain peut être faux pour un autre sans pouvoir donner impartialement la raison à l'un ou à l'autre).

S'il existait une KB regroupant les informations subjectives, contestables ou contradictoires, elle serait complémentaire des KB standard et représentative de la diversité d'opinions, de la pensée humaine globale, d'une société, d'un groupe ou d'un individu [7].

Cette ressource peut avoir un intérêt réel en plusieurs branches du Traitement Automatique des Langues, notamment en IA, en classification de données et en génération de texte.

Nous manquons les moyens en temps et en ressources pour implémenter une telle KB, mais en assimilant le mieux possible les "triplets de sortie OIE" aux faits des KB, nous pouvons faire un premier pas vers un protocole permettant d'obtenir les triplets nécessaires pour peupler une telle KB.

[7] Il n'est pas question ici de proposer une KB contenant des triplets sans structure de fait, sans discriminer les triplets par entité ou en incorporant des triplets sans filtrer leur contenu. L'idée théorique est de proposer une KB incorporant des faits tout aussi informatifs mais non vérifiés, moins factuels et plus "humains" ; une KB pouvant contenir invariablement les faits "The Matrix : is : a good movie", "The Matrix : is : an awful movie", "The Matrix : is : mediocre", "The Matrix : is : the best movie ever made".

CHAPITRE 2

RESSOURCES ET OUTILS

Notre objectif pour ce mémoire [1] est d'utiliser les outils, les corpus et les méthodes déjà existantes pour assimiler le mieux possible les triplets de sortie OIE aux faits des KB.

Pour cela, il nous faut tout d'abord nous pencher sur les outils et ressources disponibles, faire un choix des mieux adaptés à notre objectif et les analyser en profondeur.

2.1 Les bases de connaissances

Il existe plusieurs KB ayant pour objectif de répertorier des connaissances en n'excluant, à priori, aucun thème.

C'est à partir du contenu de ces KB que nous analyserons à quel point les triplets de sortie OIE offrent des faits différents aux faits habituels des KB. Il aurait été souhaitable d'extraire les données présentes dans toutes les KB à notre disponibilité. Malheureusement, comme chaque KB possède ses propres moyens de consultation, formats et structures, nous avons dû limiter notre analyse aux KB les plus complètes.

Comme nous le remarquerons dans le tableau, 2.I les KB possédant le plus grand nombre d'entités (concepts sémantiquement distincts) sont Wikidata, Freebase et le Google Knowledge Graph. DBpedia possède davantage de faits que Wikidata, ce qui assure une plus grande quantité d'information par entité, mais le nombre d'entités DBpedia représente moins d'un quart des entités Wikidata ce qui réduit considérablement la probabilité de trouver des entités spécifiques et particulières. Qui plus est, Wikidata connaît actuellement une expansion assez importante : depuis septembre 2013 jusqu'à aujourd'hui, DBpedia a incorporé environ 58 000 entités, pendant le même laps de temps, Wikidata

[1] En ayant en tête de futures recherches et la création prospective d'une KB à faits subjectifs.

a incorporé 124 074 000 entités et continue de croître de 3 millions d'entités par mois[2] approximativement.

	Nb entités (millions)	Nb faits (millions)
YAGO	10	120
NELL	2.98	90
DBpedia	4	580
Deep Dive	0.055	0.980
Wikidata	25	144
Freebase	58	1900
Google Knowledge Graph	1000	70000

Tableau 2.I : Taille approximative des différentes KB analysées

Notre choix s'est donc porté sur Wikidata, Freebase et le Google Knowledge Graph, qui, bien qu'elles se référencent mutuellement, possèdent chacune ses particularités et peuvent contenir des faits ou des entités présentées différemment ou n'apparaissant pas dans les autres KB.

2.1.1 Wikidata

Le projet Wikidata a été lancé par l'organisation Wikimedia Deutchland en 2012 afin de proposer une KB hébergée par la fondation Wikimedia (Wikipédia, Wiktionary, Wikimedia Commons, etcætera), éditée collaborativement [3] et cherchant à structurer toutes les données des autres projets Wikimedia (cf. annexe IV).

Selon Vrandečić et Denny [15] Wikidata ne se limite pas exclusivement à des faits immuables et peut contenir des faits incohérents et contradictoires afin de refléter la diversité de connaissances et la divergence d'opinions d'une entité donnée.

Cependant, ces faits subjectifs sont limités ne peuvent être que des instances de catégories permises par l'ontologie de Wikidata et leur nombre au total est minimal. Même

[2] https ://tools.wmflabs.org/wikidata-todo/stats.php
[3] Au travers de leur interface web.

en ajoutant expressément des faits subjectifs ou de nouvelles instances, cette KB est librement éditable par les utilisateurs humains, et ceux-ci peuvent être rapidement réfutés, remplacés et supprimés par la communauté.

2.1.2 Freebase

La KB Freebase a été lancée en mars 2007 par l'entreprise Metaweb. Il s'agit d'une KB éditée collaborativement [4].

Le format utilisé pour les faits Freebase est un format de triplet S-P-O analogue à RDF où chaque entité[5] était associée à une ou plusieurs catégories[6] requérant certains types de données (cf. annexe V).

En 2010, l'entreprise Metaweb a été acquise par Google et, avec elle, la KB Freebase. Google a annoncé que Freebase servirait de point de départ pour la mise en place d'une nouvelle KB destinée à prendre la relève et appelée Knowledge Graph. Malgré l'achat de la compagnie, et la mise en place en 2012 de la KB Knowledge Graph, Freebase a continué à grandir et à recevoir des contributions jusqu'en 2015 [6], date à laquelle le projet Freebase a été arrêté définitivement et la plateforme a été retirée. Malgré cela, Google a mis à la disposition du public des "dumps" bruts de Freebase dans l'état où cette KB se trouvait en 2015, ce qui permet de continuer à l'interroger et à l'exploiter.

C'est une des raisons pour laquelle les résultats Freebase pour une requête donnée sont souvent réduits, obsolètes ou inexistants : car l'ajout d'entités et de données date de 2015 et le temps a manqué pour compléter tous les faits sur certaines entités moins connues du grand public (ayant donc, moins de contributeurs). Toutefois, la quantité d'entités et de faits que Freebase a réussi à accumuler pendant ses huit années d'existence sont loin d'être négligeables et il nous a semblé pertinent de prendre cette KB en compte.

[4] Au travers de l'interface mise en place par Metaweb, le public de contributeurs non spécialisés était libre de modifier les informations générales et les métadonnées de chaque entité, ainsi que d'ajouter ou supprimer les types de données associées à chaque catégorie (pour plus de détails cf. O'Reilly [11]

[5] "Topics" dans la terminologie Freebase.

[6] "Types" dans la terminologie Freebase, une sorte de gabarit ("template").

2.1.3 Knowledge Graph

Knowledge Graph est une KB développée par Google et mise en place depuis 2012 afin de complémenter leur moteur de recherche. L'idée étant que si l'utilisateur n'est intéressé que par les informations principales de la requête, le Knowledge Graph peut fournir ces informations sous forme de tableau dans la "marge" des résultats de recherche. Cette fonctionnalité a pour effet de requérir toujours autant l'utilisation du site Google, mais il réduit considérablement les consultations des sites où, préalablement, l'utilisateur pouvait trouver ces informations.

Selon Google, Knowledge Graph utilise plusieurs sources pour puiser l'information requise. Parmi ces sources nous pouvons citer Freebase, Wikidata, Wikipédia, le CIA World Factbook et IMDB. Alors que Wikidata et Freebase étaient librement modifiables et consultables dans leur totalité, le Knowledge Graph est un produit appartenant à Google et nous conjecturons qu'en tant que tel il n'est disponible au public général ou spécialisé que dans une version réduite ou épurée[7].

Le Knowledge Graph présente une quantité réduite de faits pour chaque entité (par rapport aux autres KB), mais le nombre d'entités consultables est si grand (cf. tableau 2.I) que nous ne pouvions négliger son inclusion dans notre projet.

2.1.4 Évaluation comparative des bases de connaissances

Avant de pouvoir comparer les faits contenus dans les KB avec les faits produits à partir d'information libre, nous devons nous interroger sur la quantité de faits déjà existante dans les KB.

Afin d'effectuer cette évaluation comparative des KB sélectionnées, nous avons utilisé une liste de titres d'articles Wikipédia en anglais sur laquelle nous avons choisi au

[7]Nous arrivons à cette conjecture en observant que les faits montrés au public dans la marge de Google Search et les faits disponibles dans leur API divergent quelquefois. Qui plus est, en analysant les informations disponibles au public dans Knowledge Graph et dans les diverses sources d'où il puise ses données (cf. annexe VI).

hasard (ou presque [8]) une centaine de titres. Nous avons utilisé ces cent titres comme entités de requête aux trois KB (qui sont, toutes trois, partiellement basées sur Wikipédia).

Cette méthode a pour point fort d'éviter les biais de préférence de choix d'entités que nous aurions rencontrés si nous-mêmes ou des participants avions choisi les entités de requête [9]. Mais elle n'est pas pour autant dépourvue de défauts : l'un des principaux étant que, quelquefois, les titres d'articles Wikipédia ne correspondent pas à une entité d'intérêt pour les KB (des titres faussés ou de faux articles, des titres d'articles de désambiguïsation, etcætera) ou bien ils n'apparaissent pas dans les bases de données (articles trop récents, articles en désaccord avec le protocole de population des KB, etcætera).

Comme notre objectif n'est pas d'évaluer la capacité des KB à inclure une entité basée sur un titre Wikipédia, nous avons rejeté de notre évaluation comparative toute entité ambiguë et toute entité n'étant pas présente dans les trois KB. Ce qui nous a donné un échantillon d'évaluation composé de vingt entités présentes dans Freebase, Wikidata et le Knowledge Graph [10].

Nous avons utilisé cet échantillon [11] pour lancer des requêtes sur les trois KB et récupérer leurs faits. Afin de juger de l'ampleur informative des KB, pour l'instant, nous nous intéressons exclusivement à mesurer la quantité de faits sans tenter de savoir combien de faits d'une KB se retrouvent dans une autre, jusqu'à quel point les faits diffèrent et combien en sont exclusifs à telle ou telle KB.

Nous analysons les données résultantes du nombre de faits en les regroupant par KB

[8] Nous nous sommes permis d'inclure parmi les cent titres d'articles une requête fétiche depuis des années au laboratoire du RALI : l'entité "Chilly Gonzales". Comme celle-ci était présente dans les trois KB choisies, elle a fini dans la sélection finale des entités choisies.

[9] Ibid.

[10] Les vingt entités étant : 'Paramecus', 'Copelatus unguicularis', '1927 World Snooker Championship', 'Moustached brush finch', 'Hans Jacob Horst', 'Easingwold', 'Hugh Borton', 'YIT', 'Battle of Whitestone Hill', 'Micropogonias', '27056 Ginoloria', 'Spargaloma sexpunctata', 'Surinder Vasal', 'MILGEM project', 'Neadeloides', 'Temnora sardanus', 'Dobrogea Veche', 'Soltam M-71', 'Chilly Gonzales', 'Hellboy : Blood and Iron'.

[11] Dans un souci de clarté et comme nous nous référerons souvent à ces vingt entités comme des requêtes d'entrée de KB ou comme thème de textes d'entrée des OIE, nous nous référerons à chacune d'entre elles sous le terme E (à ne pas confondre avec le terme "entité" se référant au concept d'entité pas à l'entité fournie comme entrée de KB ou OIE).

(figure 2.1), par E [12] (figure 2.3) et par l'ensemble d'E et de KB (figure 2.5). Nous avons également réalisé des graphiques de boîtes à moustaches afin de pouvoir observer dans un même graphique la moyenne, la médiane et les extrêmes. Dans ces graphiques la quantité de faits est analysée selon la KB dont elle est extraite (figure 2.2) et selon l'E dont elle est extraite (figure 2.4).

Après une analyse superficielle de cette évaluation comparative nous observons que :

- En nous basant sur la figure 2.1, nous remarquons que Freebase est la KB avec le plus grand nombre de faits pour l'échantillon des E. Cependant, sur les figures 2.2 et 2.5, nous remarquons que ceci est dû à un cas isolé et, si nous nous fions à la médiane (sur la figure 2.2), Wikidata est la KB qui a le plus grand nombre de faits par E de façon constante.

- En observant les figures 2.2, 2.2 et 2.5 nous remarquons que Knowledge Graph est la KB avec le moins de faits, mais c'est également la KB le nombre de faits le plus constant (chaque E contient entre 4 et 10 faits).

- La figure 2.4 indique un nombre de faits par E constant, ayant une médiane tournant autour de 10 faits sauf dans le cas de deux E que nous pouvons classifier comme plus "populaires" et donc plus enclins à être édités et enrichis de faits par la communauté.

2.2 Outils : Les extracteurs d'information libre

Lorsque nous analysons les faits comme unités minimales des KB, nous observons que leur forme correspond à (ou est dérivée de) un triplet sémantique S-P-O [13].

Si nous désirons comparer le contenu informatif des différentes KB et le contenu informatif de textes libres et non structurés, il nous faut transformer le texte libre en

[12] Entités.

[13] À la vue de l'utilisateur, certaines KB ne montrent pas directement la forme S-P-O et omettent ou prennent pour acquis S [14]. Cette omission a pour but d'éviter la répétition inutile du S.

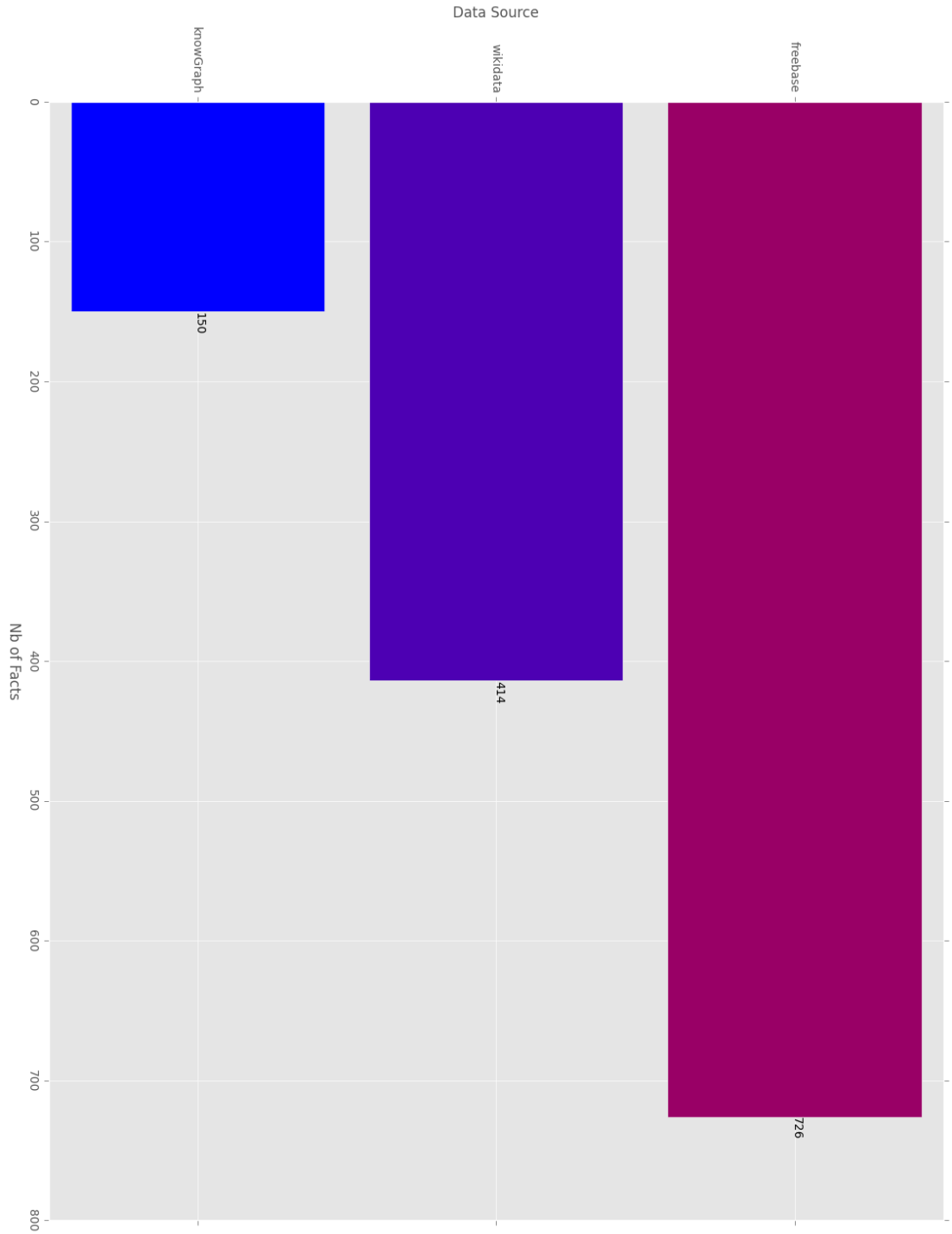


Figure 2.1 : Nombre de faits de l'échantillon évalué, segmenté par KB, toutes E confondues.

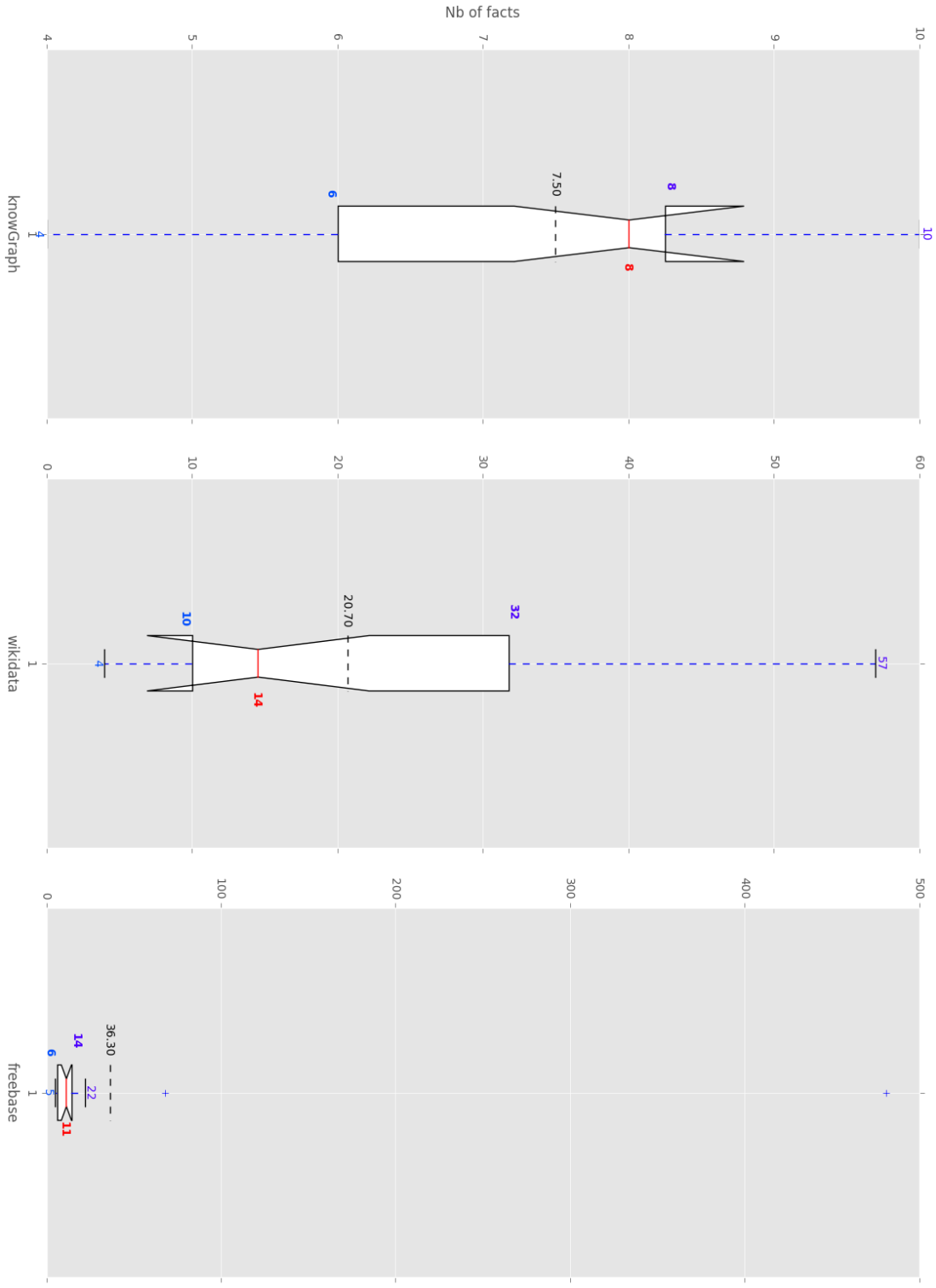


Figure 2.2 : Boîtes à moustaches des faits de l'échantillon évalué, segmenté par KB, toutes E confondues. En rouge : la médiane, en noir : la moyenne, en bleu et en violet : le maximum et minimum des moustaches.

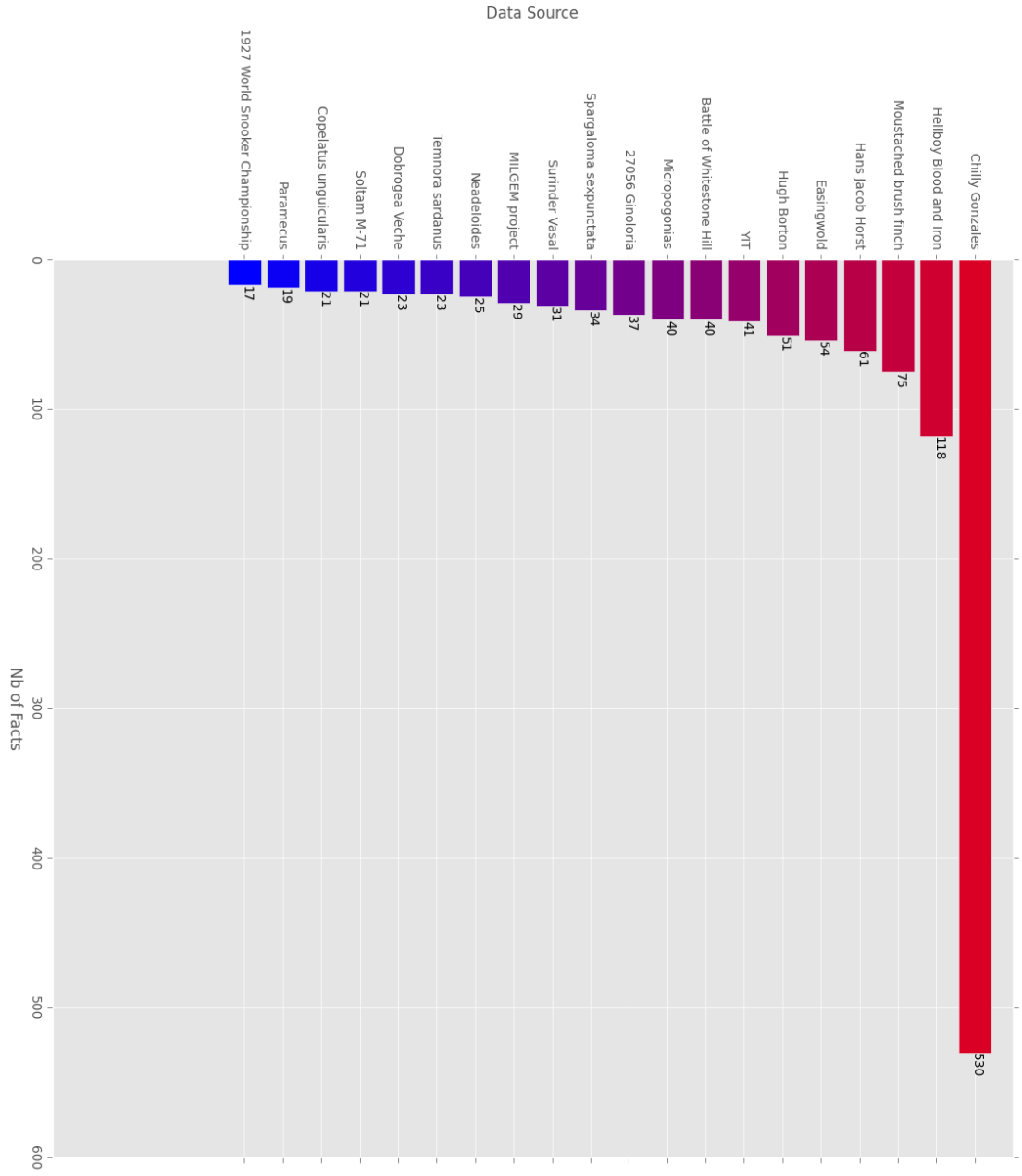


Figure 2.3 : Nombre de faits de l'échantillon évalué, segmenté par E, toutes KB confondues.

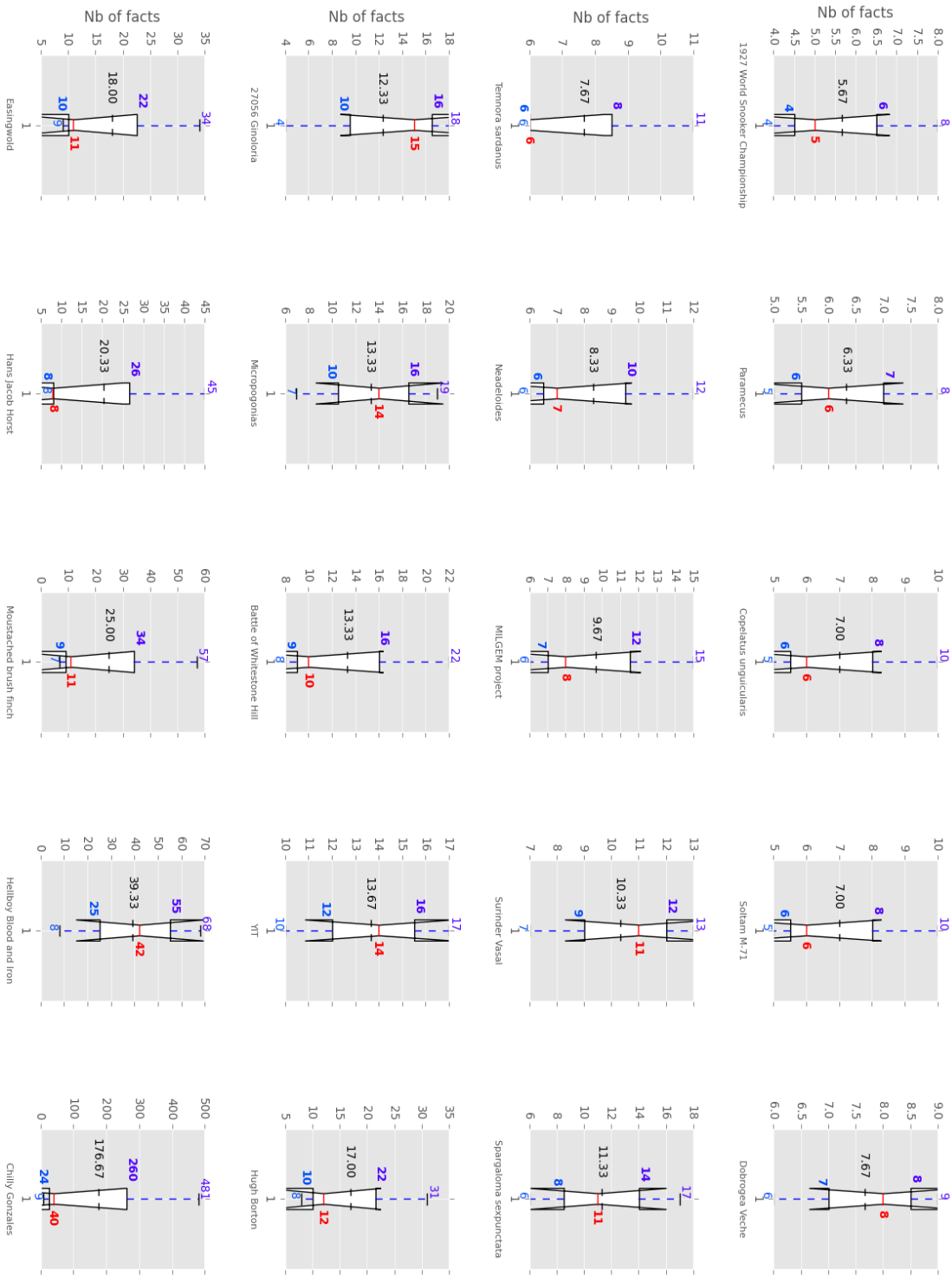


Figure 2.4 : Boîtes à moustaches des faits de l'échantillon évalué, segmenté par E, toutes KB confondues. En rouge : la médiane, en noir : la moyenne, en bleu et en violet : le maximum et minimum des moustaches.

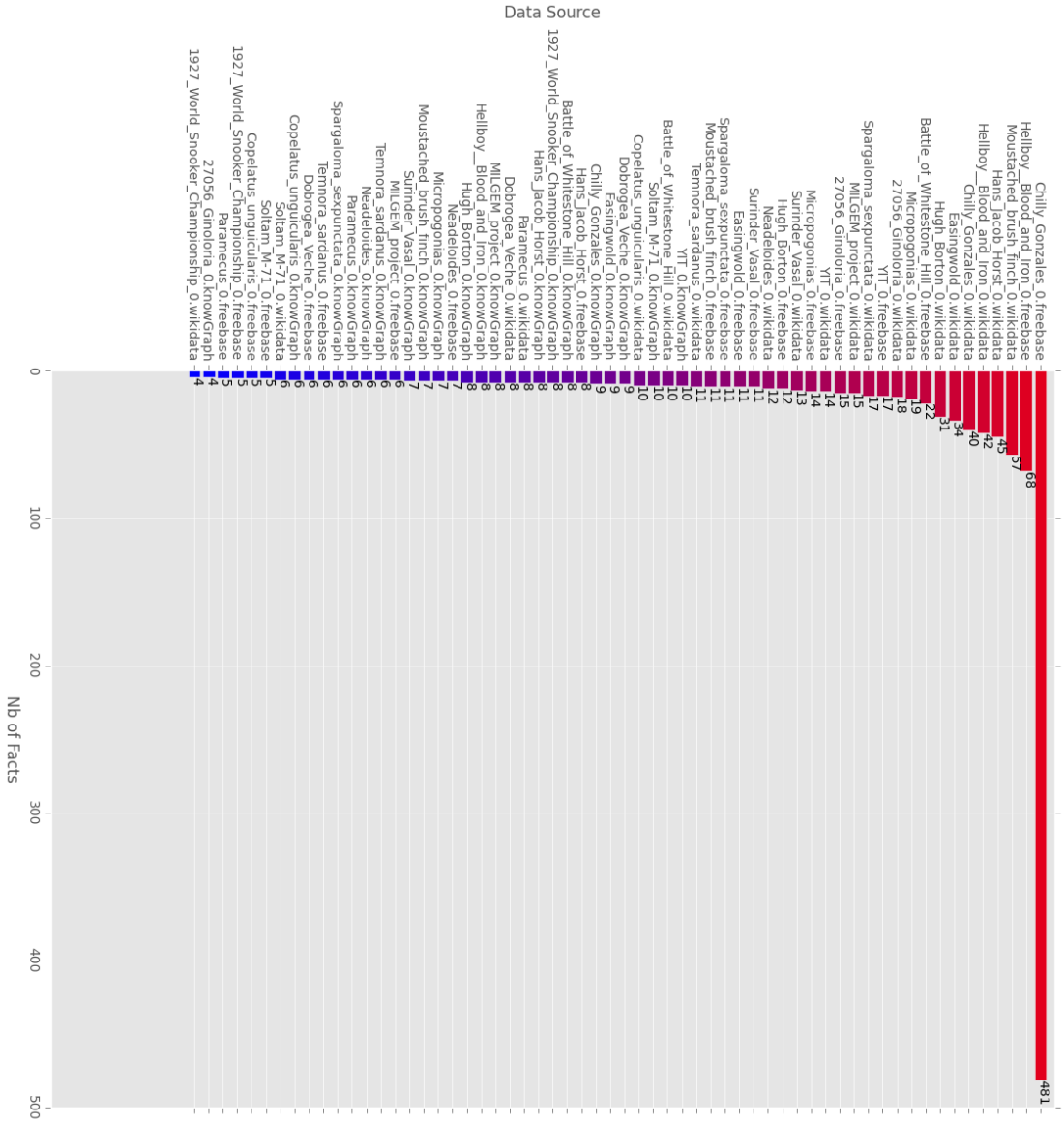


Figure 2.5 : Nombre de faits de l'échantillon évalué, segmenté par E et par KB.

un format d'unités minimales identique ou similaire aux unités minimales des KB : les triplets S-P-O. Pour ce faire il faut prendre en compte plusieurs aspects langagiers particuliers comme la syntaxe, la morphosyntaxe et la sémantique. Fort heureusement, depuis 2007 nous observons un engouement pour le développement d'OIE.

Ces systèmes nous permettent d'avoir en entrée du texte libre et d'obtenir en sortie des triplets S-P-O (plus ou moins) correspondants aux différentes unités sémantiques minimales du texte.

Comme nous pouvons le remarquer, sur ce point, les triplets de sortie OIE et les faits de KB correspondent, mais outre leur forme S-P-O, nous pourrions nous demander s'il existe d'autres points communs ou divergences.

L'objectif final des KB et des OIE est le même : fournir de l'information sur un thème spécifique sous forme de triplets S-P-O. Bien sûr, l'entrée nécessaire et la sortie de chaque type d'outil diffèrent : pour une KB il faut comme entrée une E et le S de la sortie correspond à 100% avec E, tandis que pour l'OIE il faut comme entrée du texte libre (information libre) ayant E pour thème global et la sortie a un lien indéniable avec le texte d'entrée, mais le lien avec E n'est pas toujours évident.

Cependant, dû à la nature des textes libres en entrée, les sorties des OIE ont certaines caractéristiques que les KB ne possèdent pas. L'ensemble des faits des KB se limitent à des informations factuelles (ce qui, au-delà de limiter leur nombre, leur donne un 'aspect' artificiel) alors que l'ensemble des triplets de sortie OIE ont un 'aspect' plus naturel puisqu'ils représentent des informations appréciatives, historiques, comparatives, etcætera.

2.2.1 Choix des systèmes extracteurs d'information libre

Le premier obstacle que nous rencontrons est de choisir les OIE les plus prometteurs parmi les OIE disponibles (KnowItAll, TextRunner, WOE, ReVerb, OLLIE, KrakeN, ClausIE, PropS, CSD-IE, OpenIE-4, Standford Open IE [CoreNLP]).

Après avoir jugé de la qualité des sorties de chacun de OIE, nous aurions pu nous

limiter à utiliser celui rendant le meilleur score, mais au vu du manque d'outils hautement fiables pour juger spécifiquement lequel est le mieux adapté à la tâche d'extraction de notre projet, nous avons décidé d'en choisir trois. Ceci nous permet de juger de prime abord lequel ou lesquels s'adaptent le mieux à nos tâches.

Pour pouvoir faire un choix justifié des trois meilleurs OIE, nous nous sommes basés en partie sur l'article de Stanovsky et Dagan (2016) [13] et nous avons utilisé leur banc d'essai (disponible sur Github [15]). Dans cet article, les chercheurs décrivent un banc d'essai, soit : un test d'évaluation des performances des OIE en matière d'analyse des scores de Rappel, de Précision et d'AUC (Aire Sous la Courbe) (cf. annexes I et II). Ils ont utilisé le corpus QA-SRL de He et Lewis et Zettlemoyer (2015) de l'Université de Washington [7] pour créer un 'Gold Standard' de 10 359 extractions sous forme de triplets S-P-O [16] qu'ils ont utilisé pour évaluer six OIE "proéminents" [17]. Dans le cadre de leur banc d'essai, Stanovsky et Dagan ne s'intéressent qu'aux triplets dont les relations sont introduites par des verbes [18]

Afin d'avoir une idée concise et étalonnée de la précision et du rappel des triplets de sortie des différents OIE, nous proposons le tableau 2.II .

Après avoir analysé ces scores, nous avons choisi trois OIE à tester pour notre projet : OLLIE [9], ClausIE [4] et OpenIE-4 [10].

Malgré ses bons scores, nous n'avons pas inclus PropS parmi nos choix d'OIE pour deux raisons :

- La raison principale pour exclure cette OIE est qu'il a des difficultés à traiter des

[15] https ://github.com/gabrielStanovsky/oie-benchmark

[16] La version du banc d'essai disponible sur GitHub comporte 8 479 extractions. Selon les propres dires des auteurs, ils ont supprimé du corpus 'Gold Standard' les triplets dont l'un des arguments était un pronom si ce triplet existait déjà dans une version ayant remplacé le pronom par une entité non-pronominale (cf. annexe III).

[17] Leur méthode de référence n'accepte que quelques formats de sortie OIE, mais avec quelques menus changements dans la présentation des triplets de sortie, il est possible d'évaluer virtuellement n'importe quel autre OIE.

[18] Pour une phrase comme "Chilly Gonzales, Canadian musician, is best known for his first MC and electro albums." le banc d'essai ne détectera pas le triplet "Chilly Gonzales : [is] : musician" comme un triplet correct car sa relation n'est pas produite par un verbe de la phrase, mais elle est inférée par la syntaxe des éléments nominaux.

	Precision	*Rappel*
OpenIE-4	0.63	0.55
ClausIE	0.50	0.74
OLLIE	0.46	0.34
PropS	0.64	0.50
Stanford	0.12	0.19
ReVerb	0.01	0.002

Tableau 2.II : Tableau rapportant la précision et le rappel de la sortie (tous scores de confiance confondus) des différents OIE du banc d'essai de Stanovsky et Dagan [13]. Scores compris en 0.0 et 1.0.

caractères en dehors de la norme ASCII (ces types de caractères sont communs pour certaines requêtes "exotiques" ou techniques). Par exemple, les triplets de sortie OIE pour l'E "François Villon" ou "π-calculus" seraient grandement limités en nombre et nous ne pourrions pas avoir une évaluation juste de nos heuristiques par S.

- À ce jour, il a fait moins l'objet de vérifications formelles par d'autres chercheurs.

Pour ce qui est des OIE que nous avons choisi, nous présentons rapidement ci-dessous quelques détails et particularités importantes.

2.2.1.1 Nota : ReVerb

Bien que ReVerb n'ait pas été inclus comme OIE à tester pour notre projet, il nous a semblé pertinent de le présenter afin de mieux comprendre le travail de Stanovsky et Dagan [13] et quelques particularités de leur méthode de référence.

Le Centre Turing de l'Université de Washington est l'un des pionniers dans le développement de systèmes d'OIE. Ils ont développé KnowItAll en 2004, TextRunner en 2007 [17], WOE en 2010 [16] et ReVerb en 2011.

Selon Stanovsky et Dagan [13], nous observons que les résultats produits par ReVerb possèdent les scores les plus bas parmi tous les OIE testés que ce soit en Précision,

Rappel ou AUC de Précision et Rappel alors qu'à première vue les sorties de ReVerb semblent, bien que bruitées, correspondre assez bien au contenu sémantique humainement interprétable [19]. Nous formulons l'hypothèse que les mauvais scores de ReVerb sont dus à l'absence totale de traitement des coréférences et à l'extraction exclusive de relations verbales explicites. Si bien d'autres OIE tentent de résoudre (du moins au niveau intra-phrase) les coréférences, quitte à produire plusieurs tentatives de triplets pour une même relation sémantique ; ReVerb, lui, n'offre aucune résolution des coréférences.

Or, le corpus QA-SRL de He et Lewis et Zettlemoyer [7] a pour règle de résoudre la corréférence pronominale. Le pronom n'est maintenu dans le triplet que dans certains cas où il existe un doute sémantique quant au référent exact.

Ce qui pourrait expliquer (du moins en partie) qu'évalués sur un 'Gold Standard' n'ayant pas de coréférents pronominaux, les triplets de sortie de ReVerb rendent un si mauvais score.

2.2.1.2 OLLIE

Le projet "Open Language Learning for Information Extraction" (OLLIE) [9] a été développé au Centre Turing de l'Université de Washington en 2012. Par rapport aux précédents OIE du Centre Turing, OLLIE augmente la portée de l'analyseur de dépendances en identifiant des noms et des adjectifs comme P [20] des triplets S-P-O (cf. figure 2.2.1.2 et note [21]) et en introduisant une étape d'analyse du contexte intervenant dans la production de triplets.

[19] Pour plus de détails sur le terme 'humainement interprétable' cf. le chapitre 3 de ce mémoire, section 'Description de la chaîne de travail', sous-section 'Première tâche : les faits KB'

[20] Propriété du triplet/fait.

[21] Le texte étant : "Chilly Gonzales, a famous Canadian musician, resided in Paris, France for several years and now lives in Cologne, Germany. Though best known for his first MC and electro albums, he is also a pianist, producer, and songwriter."

```
Chilly Gonzales, a famous Canadian musician, resided in
Paris, France for several years and now lives in Cologne,
Germany.
0.874: (Chilly Gonzales; resided in; Paris)
0.855: (Chilly Gonzales; resided for; several years)
0.695: (Chilly Gonzales; now lives in; Cologne)

Though best known for his first MC and electro albums,
he is also a pianist, producer, and songwriter.
0.652: (he; is also; a pianist , producer , and songwriter)
```

Figure 2.6 : Sortie de OLLIE pour un extrait de texte ayant pour thème "Chilly Gonzales" (cf. note au pied de la page 21). Sous la phrase d'origine, nous observons que chaque ligne montre le score de fiabilité suivi du triplet sous la forme : (S ; P ; O).

2.2.1.3 ClausIE

ClausIE est un OIE développé en 2013 par l'Institut d'Informatique Max Planck à Saarbrücken. Cet OIE propose une nouvelle approche centrée sur les propositions grammaticales [22] qui composent chaque phrase (cf. figure 2.2.1.3).

2.2.1.4 OpenIE-4

Le projet OpenIE-4 [23] a été développé par le Centre Turing de l'Université de Washington.

Comme son nom l'indique, OpenIE-4 [10] est la quatrième version du projet. La première est parue sous le nom de TextRunner, la deuxième sous le nom de ReVerb et la troisième sous le nom de OLLIE. Cette quatrième version diffère des versions précédentes autrement que par l'originalité du nom : OpenIE-4 est en fait la combinaison de RelNoun [12] et SrlIE, deux OIE développés indépendamment.

RelNoun est un OIE spécialement conçu pour l'extraction des triplets à relation no-

[22] "Clause" dans leur terminologie, définie en surface par Del Corro et Gemulla [4] comme une part de la phrase exprimant un fragment d'information cohérent.

[23] Il nous faut faire attention de ne pas confondre OpenIE-4 (https ://github.com/allenai/openie-standalone) et le Stanford OpenIE (https ://nlp.stanford.edu/software/openie.html) qui est, quant à lui, un tout autre OIE destiné à faire partie de la boîte à outils CoreNLP de l'Université de Stanford.

```
1 "Chilly Gonzales" "is" "a famous Canadian musician"
1 "Chilly Gonzales" "resided" "in Paris France for
    several years"
1 "Chilly Gonzales" "lives" "in Cologne Germany now"
1 "Chilly Gonzales" "lives" "in Cologne Germany"
2 "his" "has" "first MC and electro albums"
2 "he" "is" "a pianist producer and songwriter Though
    best known for his first MC and electro albums"
2 "he" "is" "a pianist producer and songwriter also"
2 "he" "is" "a pianist producer and songwriter"
```

Figure 2.7 : Sortie de ClausIE pour un extrait de texte ayant pour thème "Chilly Gonzales" (cf. note au pied de la page 21). Nous pouvons lire sur chaque ligne l'index de la phrase suivi du triplet sous la forme : "S" "P" "O".

minale en mettant l'accent sur la détection de groupes de noms relationnels [24] de noms communs à majuscule [25] et de gentilés[26].

SrlIE est un OIE pour lequel nous avons très peu de détails. Nous savons qu'il est inspiré des idées proposées par Christensen et al. [3] et qu'il est basé sur un système d'étiquetage de rôle sémantique (SRL, "Semantic Role Labeling"). N'ayant aucun article scientifique ou documentation détaillée traitant SrlIE et n'ayant qu'un article traitant superficiellement OpenIE-4, il nous est difficile de comprendre le fonctionnement exact de SrlIE.

L'union de ces deux OIE en un seul permet d'obtenir des résultats extrayant tout autant les propriétés verbales que les propriétés nominales ou adjectivales (cf. figure 2.2.1.4).

[24] ie : "Google CEO Larry Page" -> (Larry Page ; [is] CEO [of] ; Google)
[25] ie : "former First Lady Michelle Obama" -> (Michelle Obama ; [is] ; former First Lady)
[26] ie : "Japanese foreign minister Kishida" -> (Kishida ; [is] foreign minister [of] ; Japan)

```
Chilly Gonzales, a famous Canadian musician, resided in
Paris, France for several years and now lives in
Cologne, Germany.
  0.86 (a famous Canadian musician; resided for; T:several years)
  0.92 (a famous Canadian musician; lives in; L:Cologne)
  0.86 (a famous Canadian musician; lives; T:now)
  0.89 (Chilly Gonzales; [is]; a famous Canadian musician)
  0.89 (Chilly Gonzales; [is] a famous musician [from]; Canada)

Though best known for his first MC and electro albums,
he is also a pianist, producer, and songwriter.
  0.45 (he; is also; a pianist, producer, and songwriter)
```

Figure 2.8 : Sortie de OpenIE-4 pour un extrait de texte ayant pour thème "Chilly Gonzales" (cf. note au pied de la page 21). Sous la phrase d'origine, nous observons que chaque ligne montre le score de fiabilité suivi du triplet sous la forme : (S ; P ; O). Nota : les éléments des Propriétés entre [crochets] marquent des inférences de l'OIE.

CHAPITRE 3

TÂCHES D'EXTRACTION

Jusqu'à présent, nous avons présenté des KB et des OIE sans vraiment approfondir sur la manière dont il est possible de les utiliser.

Dans ce chapitre nous décrirons brièvement les recherches et projets similaires au nôtre afin de présenter l'état de l'art actuel puis nous décrirons nos méthodes et nos résultats obtenus.

3.1 Projets similaires et état de l'art

Après une recherche assidue, nous n'avons pas trouvé de projet ou groupe de recherche partageant nos mêmes objectifs. Ceci est probablement dû au fait que tant les KB comme les OIE sont des types d'outils relativement récents et, bien qu'ils soient jugés particulièrement bons, ils sont encore loin d'être parfaits. Plusieurs équipes y dédient encore beaucoup d'efforts à perfectionner les différents systèmes pour obtenir de meilleurs résultats.

Certains projets comme, par exemple, YAGO [14], YAGO2 [8], DBpedia [1], NELL[2] se disent des ontologies, des bases de données ou des systèmes d'apprentissage sémantique, mais, bien que leur objectif individuel varie grandement, nous nous sentons assez confiants pour les classifier également comme des KB puisqu'ils peuvent entreposer des triplets sémantiquement corrects (obtenus en scrappant le web).

Cependant, dans un souci d'uniformité, ces systèmes associent chaque entité à une catégorie prédéfinie (en se basant sur une ontologie fixe) et ne retiennent que les triplets aux P correspondant à l'identique avec les P associées à la catégorie prédéfinie.

Prenons l'exemple d'une KB simple et fictive où l'entité "Chilly Gonzales" appartient au type "human". Ce type accepte les P "name is", "is also known as", "was born", "died", etcætera. Parmi les triplets :

- Chilly Gonzales : **name is** : Jason Charles Beck
- Chilly Gonzales : birth name is : Jason Charles Beck
- Chilly Gonzales : was born : Jason Charles Beck
- Chilly Gonzales : real name is : Jason Charles Beck

Seul le premier peut être incorporé à la KB, car c'est le seul correspondant aux P pour ce type [1].

C'est cette délimitation qui partage et distance l'objectif des populateurs de faits des KB habituelles à l'objectif de notre projet.

Bien que conventionnel dans sa façon de peupler la KB, un projet de recherche a tout de même attiré notre attention par les affirmations et résultats prometteurs déclarés : le Google Knowledge Vault.

3.1.1 Knowledge Vault

Le Knowledge Vault est une KB décrite dans l'article de Dong et al. [5]. Bien que l'article laisse entendre que le Knowledge Vault est déjà fonctionnel et actif, les représentants de Google ont précisé (en 2014) qu'il ne s'agit que d'un projet de recherche qui n'est pas encore en voie de développement [2].

L'article propose d'utiliser les données déjà existantes dans Freebase et Knowledge Graph comme ressource d'entraînement d'un modèle probabiliste. Selon Dong et al. [5] en utilisant ce modèle entraîné et de l'information libre comme entrée du Knowledge Vault, ils obtiennent une KB peuplée automatiquement sans intervention humaine, bien plus vaste (regroupant davantage d'entités) et complète (regroupant davantage de faits par entité) que toutes les KB précédentes.

[1] Et encore faut-il que l'entité soit associée au type correct ou que le(s) type(s) auquel l'entité correspond existe dans l'ontologie.

[2] http ://searchengineland.com/google-builds-next-gen-knowledge-graph-future-201640

Selon Dong et al. [5] chaque sortie du Knowledge Vault possède un score de confiance permettant de mesurer à quel point le système est certain de la véracité du fait.

Par contre, l'article ne précise pas si les triplets de sortie de l'OIE du Knowledge Vault incluent des faits subjectifs (ce qui pourrait représenter un point commun entre le Knowledge Vault et notre projet).

Si l'OIE du Knowledge Vault a été programmé pour être "inclusif" ("conservateur"), alors les triplets de sortie de cet OIE regroupent la totalité des triplets indépendamment de leur score. Et il serait possible de trouver, parmi les triplets de bas score, des faits subjectifs (en plus des triplets sémantiquement erronés).

Par contre, si l'OIE du Knowledge Vault a été programmé pour être "exclusif", alors les faits subjectifs sont exclus des triplets de sortie de cet OIE. Ce qui veut dire que le Knowledge Vault n'a pas plus de similitudes avec notre projet que n'importe quel autre projet ou groupe de recherche de KB.

Afin d'avoir plus d'information, nous aurions besoin de consulter davantage de littérature sur la Knowledge Vault (malheureusement inexistante) ou avoir accès aux faits produits (inaccessibles).

3.2 Description de la chaîne de travail de notre projet

3.2.1 Première tâche : extraction des faits KB

La première tâche consiste à extraire, pour chaque requête donnée, les faits humainement interprétables contenus dans les KB choisies, et ce dans un format uniforme. Pour cette tâche, nous nous sommes limités à l'extraction de faits humainement interprétables, car les triplets de sortie OIE se limitent évidemment à ce type spécifique de faits, et il serait donc contre-productif de comparer et d'évaluer des données de nature différentes.

Par "faits humainement interprétables" nous entendons les faits conçus pour être instinctivement lus et compris par les utilisateurs (de la KB) humains et non spécialisés. Les faits exclus sont ceux n'apportant qu'une valeur ontologico-sémantique destinée aux sys-

tèmes artificiels, classables comme "méta-données" (cf. la figure 3.1) ; cela se traduit de manière très concrète : les faits exclus sont tous les faits (ou la chaîne de faits hyperliés) dont l'objet final (aboutissement de la chaîne de faits) n'est pas une information visible dans l'interface graphique (ie : les chaînes de caractères non balisées dans la figure 3.1), mais l'instance d'une classe ontologique (ie : les chaînes de caractères entre balises dans la figure 3.1).

En analysant les faits humainement interprétables des différentes KB, nous observons que, pour une même entité d'entrée (E), les KB coïncident rarement sur les faits présentés. Pour avoir une idée plus précise, nous avons réalisé un tableau de contingence du nombre de faits communs des différentes KB (cf. tableau 3.I).

	Freebase	Knowledge Graph	Wikidata	TOTAL FAITS
Freebase	-	63	60	770
Knowledge Graph	-	-	39	250
Wikidata	-	-	-	394
TOTAL FAITS	770	250	394	

Tableau 3.I : Tableau de contingence montrant pour les différentes KB choisies le nombre de faits communs (correspondance entre le S (sujet) et l'O (objet) de chaque fait des KB choisies) aux vingt E de l'échantillon.

Pour ce tableau, nous comptabilisons l'intersection de faits lorsque le S [3] et l'O [4] d'un fait analysé ont une correspondance exacte (similitude à 100%) non sensible à la casse avec le S et l'O d'un autre fait analysé [5]. Nous n'avons pas pris en compte la P [6] du fait pour comptabiliser les faits communs, car chaque KB a sa terminologie et sa façon particulière de représenter les P (bien que le nombre de P soit limité dans chaque

[3] Sujet du triplet/fait.

[4] Objet du triplet/fait.

[5] Nous aurions pu être encore moins stricts à l'heure d'effectuer ces comparaisons en prenant en compte les divergences orthographiques et de présentation des S et des O. Nous avons choisi cette comparaison afin d'évaluer les faits partagés (avec toute l'uniformité que cela implique), pas les faits à des niveaux de similarités différents.

[6] Propriété du triplet/fait.

```
<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
                <http://rdf.freebase.com/ns/common.topic>

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
                <http://rdf.freebase.com/ns/base.type_ontology.physically_instantiable>

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
                <http://rdf.freebase.com/ns/base.type_ontology.animate>

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
                <http://rdf.freebase.com/ns/people.person>

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
                <http://rdf.freebase.com/ns/base.type_ontology.agent>

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://rdf.freebase.com/ns/type.object.name>
                "Gonzales"@fr

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://rdf.freebase.com/ns/people.person.place_of_birth>
                <http://rdf.freebase.com/ns/m.052p7>
                        <http://rdf.freebase.com/ns/type.object.name>
                                "Montréal"@fr

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://rdf.freebase.com/ns/type.object.key>
                "/wikipedia/fr/Chilly_Gonzales"

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://rdf.freebase.com/ns/common.topic.description>
                "Gonzales, ou encore Chilly Gonzales, de son vrai nom Jason Beck, est un musicien canadien. Gonzales
                produit une musique electro-pop volontairement humoristique, un « cabaret dada » avec des textes
                faussement naïfs et remplis d'autodérision, versant parfois dans une parodie de hip-hop. Il collabore
                régulièrement avec d'autres musiciens canadiens, tels que Feist, Peaches ou encore Mocky. Il a
                également collaboré avec Jamie Lidell sur ses albums Multiply et Compass, avec Buck 65 sur l'album
                Secret House Against the World et Socalled sur ses albums Ghettoblaster et Sleepover. Plus récemment,
                il a participé à l'album Random Access Memories du duo électronique Daft Punk."@fr

<http://rdf.freebase.com/ns/m.01w5ts6>
        <http://rdf.freebase.com/ns/common.topic.official_website>
                <http://www.chillygonzales.com/>
```

Figure 3.1 : Extrait de l'entité Freebase m.01w5ts6 ("Chilly Gonzales") montrant en bleu les faits de métadonnées (appartenant aux classes : thème, entité physique, entité animée, entité personne, entité de type agent, [nom d'objet])
et en rouge les faits humainement interprétables (nom, lieu de naissance, url Wikipédia simplifié, description, site officiel).

KB). Si nous analysons, à titre d'exemple, une catégorie de fait habituel comme le fait correspondant au "nom" de l'E "Chilly Gonzales", nous observons les faits suivants :

- **Freebase** → Chilly Gonzales : type.object.name : Chilly Gonzales
- **Knowledge Graph** → Chilly Gonzales : name : Chilly Gonzales
- **Wikidata** → Chilly Gonzales : label : Chilly Gonzales

Comme nous pouvons le remarquer, les faits correspondent tous trois en S et en O mais la P, bien que compréhensible dans les trois cas, diverge d'une KB à l'autre. Au-delà d'une observation purement empirique, pour en arriver à la conclusion que la P n'est pas un paramètre fiable pour mesurer l'équivalence des faits entre plusieurs KB, nous avons utilisé notre échantillon (l'ensemble de faits obtenus en interrogeant trois KB pour vingt E) pour réaliser plusieurs tableaux de contingence en comptabilisant séparément les S et P et O des faits, les S et P des faits, les S et O des faits, et finalement les P et O des faits [7]. En observant le tableau comparatif des S et P des faits, nous remarquons que les P communes sont proches de zéro, ce qui vient soutenir notre observation empirique et justifie sa non-inclusion dans notre évaluation comparative (cf. tableaux de l'annexe VII).

Pour en revenir au tableau 3.I, nous remarquons que, même en faisant abstraction des P, l'intersection des KB est très faible (entre 5 et 7 %). Ces chiffres semblent indiquer que les différentes KB contiennent des faits aux contenus très différents.

Évidemment, cette manière de mesurer les faits communs laisse échapper certains faits portant le même sens, mais présentés différemment dans l'O. Un des exemples les plus courants et notables est la structure des faits correspondant aux dates [8] :

- **Freebase** → Chilly Gonzales : people.person.date_of_birth : 1972-03-20

[7] Nous rappelons que les S dans les KB correspondent aux E et que ceux-ci correspondent en tous points dans les faits de chaque KB.

[8] Nous présentons ces faits référents aux dates à mode d'exemple des faits variant dans leur présentation, mais à l'heure de faire les statistiques présentées ici ces faits ont été uniformisés, comme spécifié à la suite de l'exemple.

- **Knowledge Graph** → Chilly Gonzales : born : March 20, 1972 (age 45), [9]
- **Wikidata** → Chilly Gonzales : date of birth.day : 20
- **Wikidata** → Chilly Gonzales : date of birth.hour : 0
- **Wikidata** → Chilly Gonzales : date of birth.minute : 0
- **Wikidata** → Chilly Gonzales : date of birth.month : 3
- **Wikidata** → Chilly Gonzales : date of birth.year : 1972

Dû à la fréquence avec laquelle nous retrouvons la structure des faits de date dans les KB, nous avons choisi d'uniformiser du mieux possible [10] les sorties des différentes KB pour obtenir un seul et même format numérique en sortie : année-mois-jour. Ceci nous permet de réaliser une meilleure analyse comparative entre les diverses KB et de ne pas laisser échapper ces faits. Il reste encore d'autres faits qui sont difficilement identifiables et normalisables. Cependant, d'après nos observations empiriques, ils ne sont pas suffisamment nombreux pour altérer considérablement notre analyse comparative des faits de différentes KB.

3.2.2 Deuxième tâche : extraction des triplets de sortie OIE

La deuxième tâche consiste à obtenir les triplets de sortie OIE dérivés de textes d'information libre ayant E pour thème.

Pour ce faire, nous utilisons E pour faire une requête à Wikipédia et tenter de trouver un article qui lui correspond. En utilisant l'URL de cet article, nous extrayons son contenu nous l'utilisons comme entrée pour chacun des systèmes d'OIE choisis.

À partir de E, nous interrogeons Google Search et nous faisons l'extraction du contenu des dix premières pages web (en excluant la page Wikipédia correspondante et les pages

[9] Non-accès à ce fait depuis l'API du Knowledge Graph de Google, fait obtenu en faisant du web scrapping du tableau Knowledge Graph de la page Google Search ayant pour requête E.

[10] Exception faite des dates approximatives, des dates inconnues, des dates d'avant notre ère, etcætera.

web dans une autre langue que l'anglais), que nous utilisons également comme entrée des trois OIE choisies.

Finalement, nous traitons les triplets de sortie OIE afin de rendre leur format uniforme en tentant de les modifier le moins possible [11].

Au final, nous obtenons pour chaque requête deux groupes uniformes de triplets de sortie OIE : un provenant de l'article Wikipédia, l'autre provenant des dix premières pages rendues par Google Search (cf. figure 3.2).

Les triplets obtenus grâce aux OIE sont plus nombreux et portent des informations plus variées, car ils correspondent aux extractions faites sur des textes libres. D'ailleurs, les textes libres thématiques offrent des informations tellement différentes des informations des faits de KB que ce n'est que très rarement que les triplets de sortie OIE se retrouvent également dans les KB (cf. tableau 3.II).

	Freebase	Knowledge Graph	Wikidata	TOTAL
ClausIE	2	0	2	9254
OLLIE	1	1	2	7981
OpenIE-4	0	0	0	3249
TOTAL	770	250	394	

Tableau 3.II : Tableau de contingence montrant le nombre de triplets communs aux KB et aux OIE choisis (toutes sources OIE confondues : corpus d'entrée tiré de Wikipédia et corpus d'entrée tiré des dix premières pages suggérées par Google Search).

Remarquons que les triplets de sortie OIE peuvent varier tellement que, bien souvent, pour un même corpus d'entrée, des OIE différents vont coïncider peu souvent (cf. 3.III).

Après la tâche décrite, en ayant obtenu des triplets de sortie OIE, nous pouvons passer à la troisième phase de filtrage en vue d'assimiler les triplets de sortie OIE aux faits de KB.

[11] À ce stade, nous les rendons uniformes en modifiant exclusivement leur format de présentation en sortie (pas leur contenu) ou en supprimant les éléments qui ne peuvent pas s'adapter au format souhaité. Nous supprimons notamment les duplets, quadruplets et autres N-plets n'étant pas des triplets.

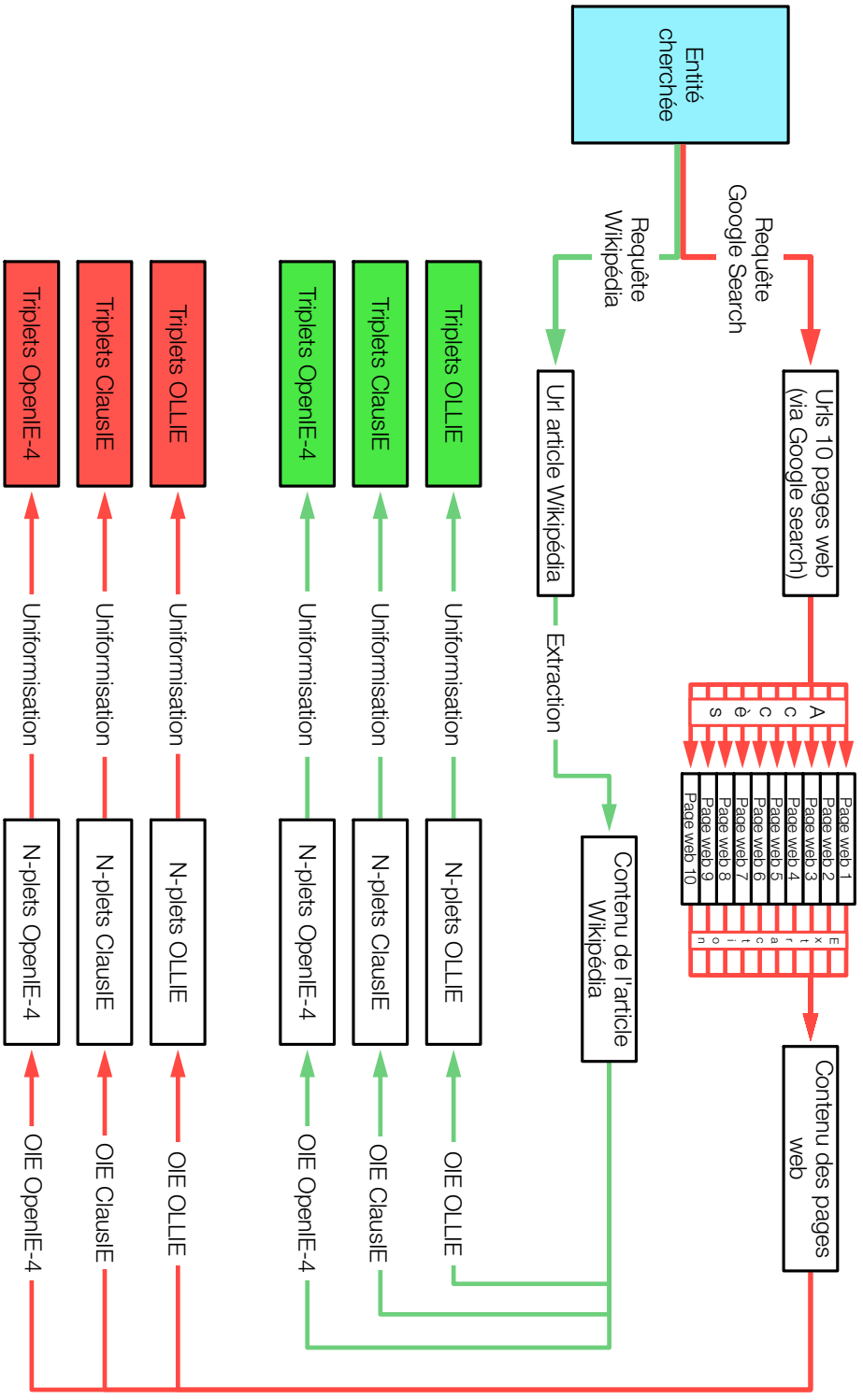


Figure 3.2 : Étapes suivies pour la tâche d'OIE sur un article Wikipédia et les pages web suggérées par Google Search ; permettant d'obtenir respectivement des triplets S-P-O uniformes en sortie de OLLIE, ClausIE et OpenIE-4.

	ClausIE	OLLIE	OpenIE-4	TOTAL
CLausIE	-	351	506	9254
OLLIE	-	-	223	7981
OpenIE-4	-	-	-	3249
TOTAL	9254	7981	3249	

Tableau 3.III : Tableau de contingence montrant le nombre de triplets communs aux OIE choisis (toutes sources OIE confondues : corpus d'entrée tiré de Wikipédia et corpus d'entrée tiré des dix premières pages suggérées par Google Search).

CHAPITRE 4

TÂCHE DE FILTRAGE ET ANALYSE

4.1 Tâche de filtrage

La tâche de filtrage consiste à réduire le bruit des triplets de sortie OIE en rejetant, en même temps, les triplets sémantiquement non pertinents et les triplets aux caractéristiques non similaires des faits de KB.

Nous aurions pu nous limiter à utiliser l'OIE ayant un meilleur score de Rappel, Précision ou F-mesure sans utiliser d'heuristiques ; mais nous rencontrons deux obstacles principaux :

- le score de Précision pour les triplets de sortie OIE le plus efficace (selon le banc d'essai de Stanovsky et Dagan [13]) n'atteint pas 0.8 (au seuil de confiance le plus bas)
- et les triplets de sortie OIE ne se limitent pas à extraire des triplets dont le S correspond à E (comme c'est le cas dans les faits de KB).

C'est pourquoi si nous voulons assimiler au mieux les triplets de sortie OIE aux faits KB, il nous faut nettoyer automatiquement les triplets de sortie OIE pour en augmenter la précision au maximum.

Nous pouvons observer que, par rapport au nombre de faits des KB (cf. figures 2.2 et 2.4 dans le chapitre *Ressources et outils*), les triplets extraits à partir d'information libre représentent une ressource bien plus abondante (cf. figures 4.1 et 4.2 obtenues à partir de l'article Wikipédia et figures 4.3 et 4.4 obtenues à partir des pages suggérées par Google Search [1]).

C'est pourquoi, si nous désirons obtenir les triplets de sortie OIE avec le meilleur score de précision possible et se rapprochant le plus possible des faits de KB (quitte à

[1] Pour les schémas sur les nombres de faits voir annexes XI, XII, XIII

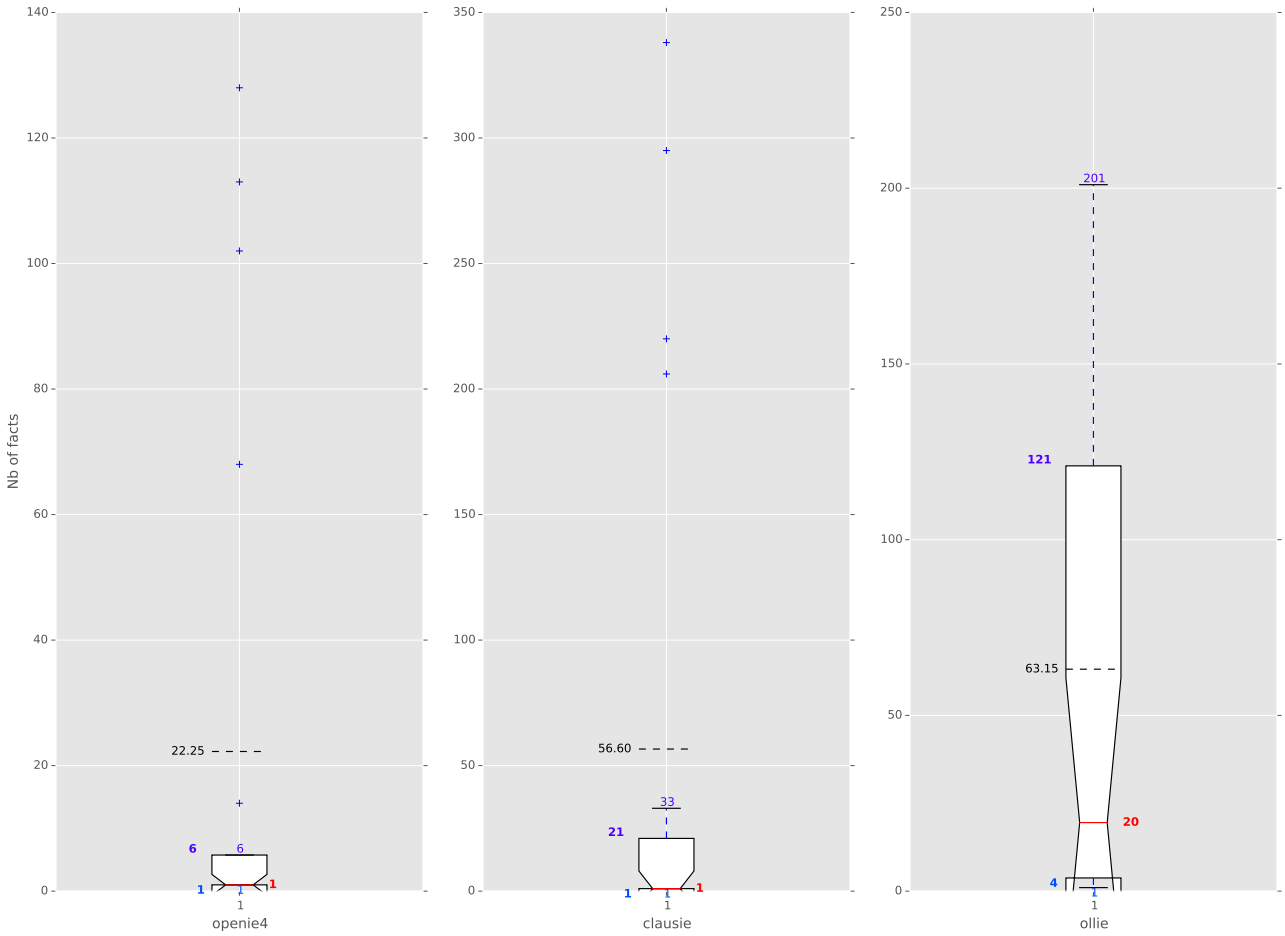


Figure 4.1 : Boîtes à moustaches des triplets de sortie OIE extraits du contenu de l'article Wikipédia. Les triplets sont segmentés par OIE, toutes entités confondues. En rouge : la médiane, en noir : la moyenne, en bleu et en violet : le maximum et minimum des moustaches.

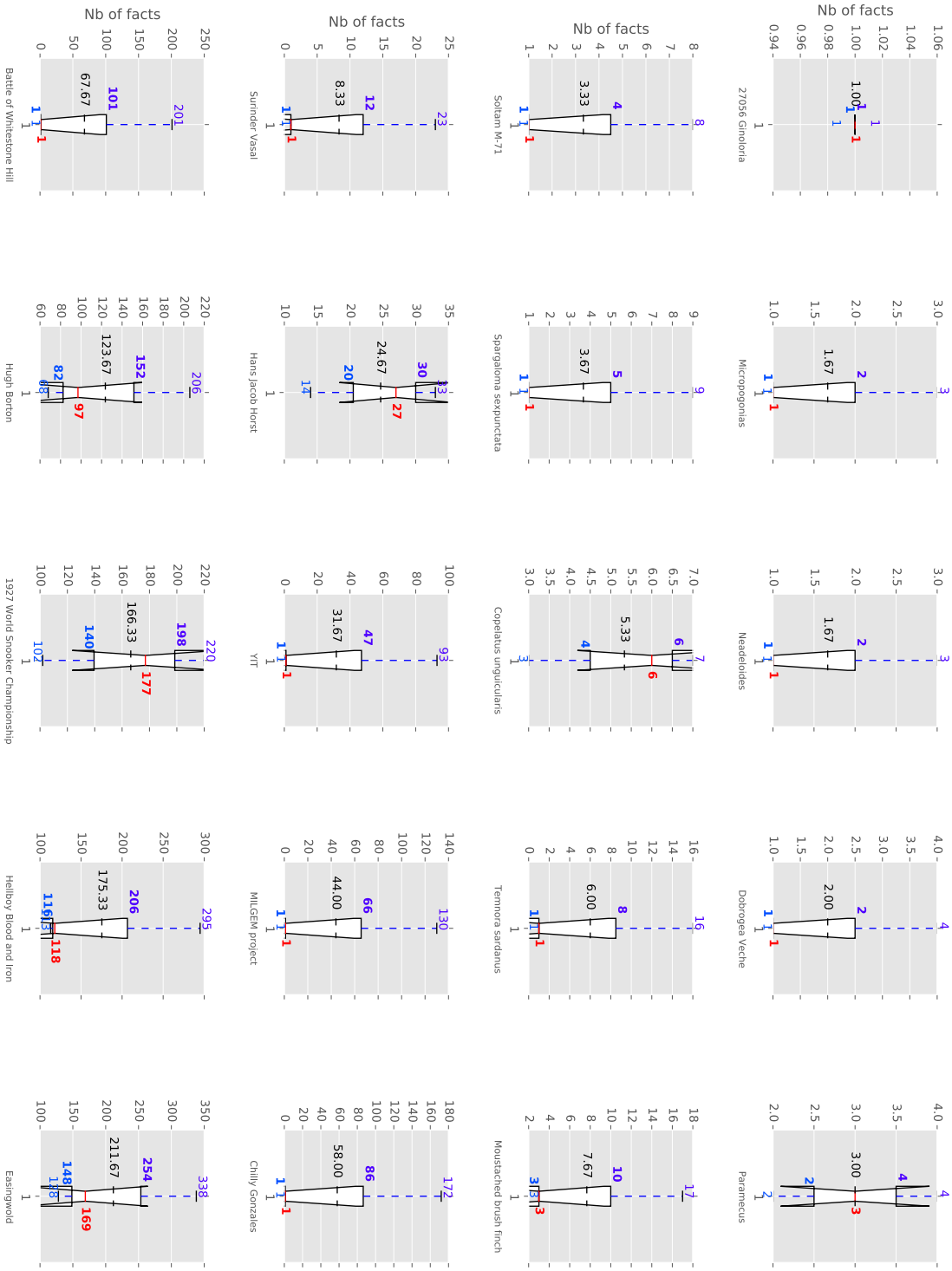


Figure 4.2 : Boîtes à moustaches des triplets de sortie OIE extraits du contenu de l'article Wikipédia. Les triplets sont segmentés par entités, tous OIE confondus. En rouge : la médiane, en noir : la moyenne, en bleu et en violet : le maximum et minimum des moustaches.

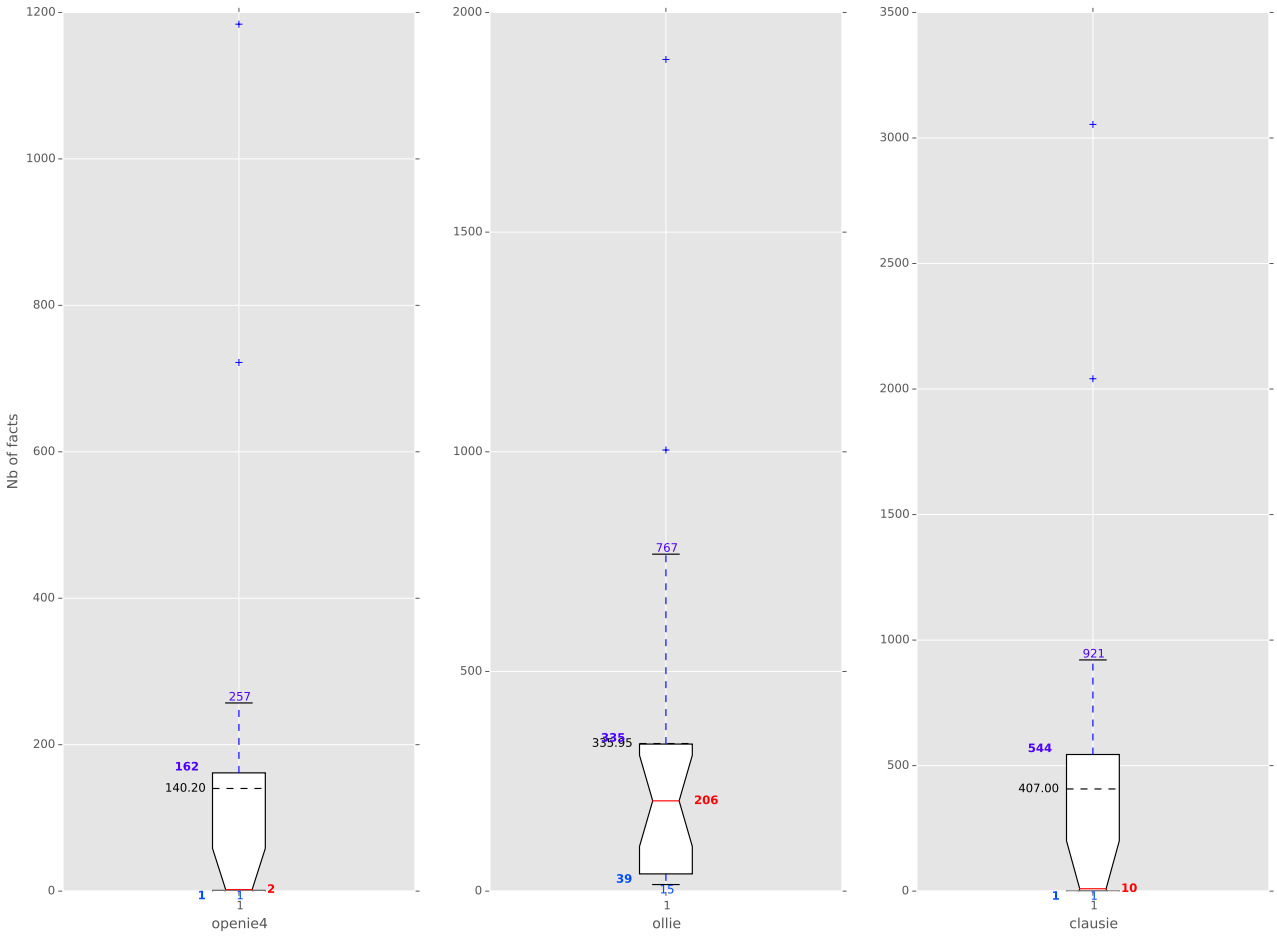


Figure 4.3 : Boîte à moustaches des triplets de sortie OIE extraits du contenu des 10 premières pages rendues par Google Search. Les triplets sont segmentés par OIE, toutes entités confondues. En rouge : la médiane, en noir : la moyenne, en bleu et en violet : le maximum et minimum des moustaches.

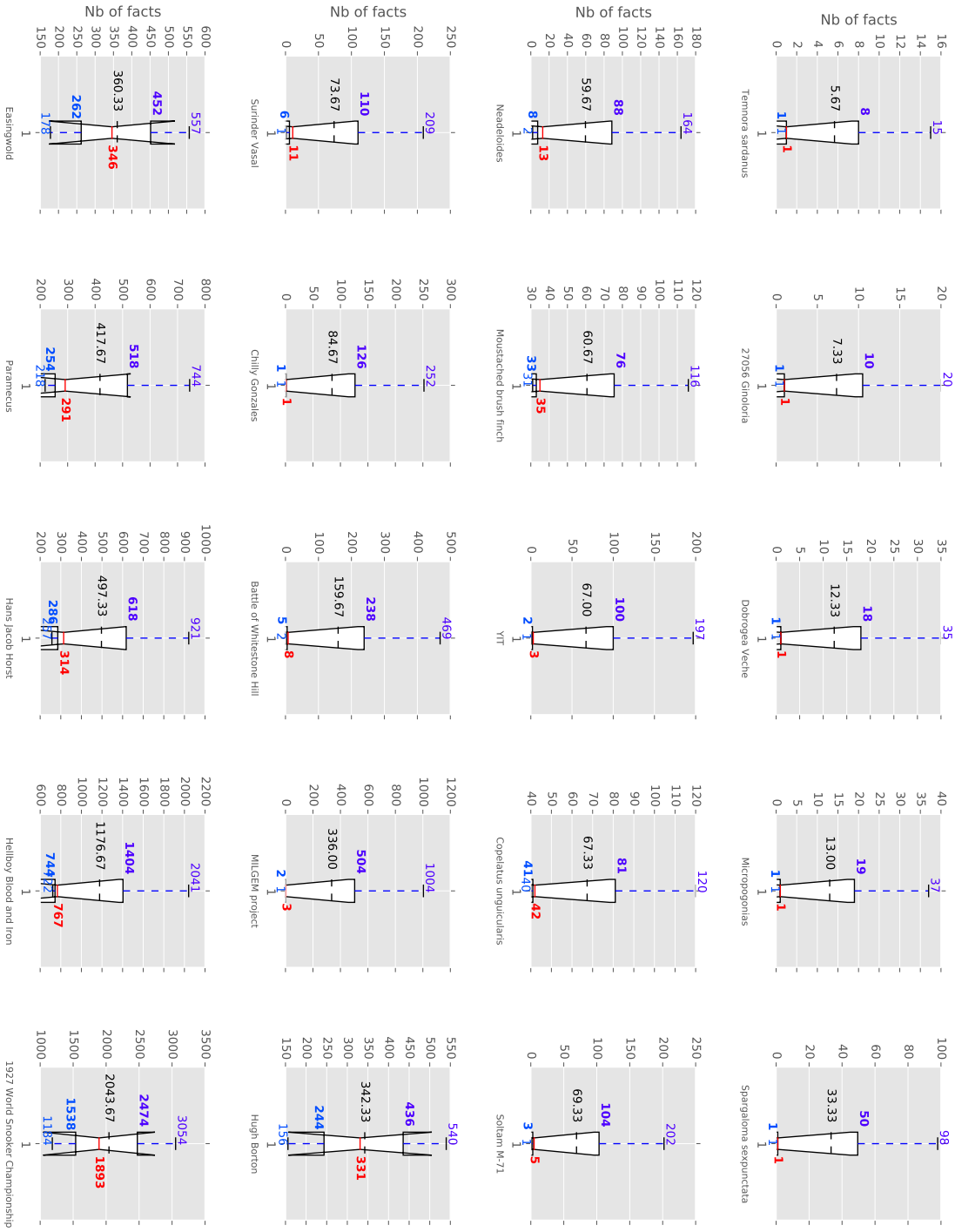


Figure 4.4 : Boîte à moustaches des triplets de sortie OIE extraits du contenu des 10 premières pages rendues par Google Search. Les triplets sont segmentés par entités, tous OIE confondus. En rouge : la médiane, en noir : la moyenne, en bleu et en violet : le maximum et minimum des moustaches.

avoir un score de rappel très bas) nous pouvons utiliser des heuristiques permettant de ne capturer que les triplets qui nous sont d'intérêt.

Afin de mieux cerner les capacités des heuristiques, dans cette section nous reprenons les données d'exemple des triplets de vingt entités choisies au hasard. Ceci nous permettra d'analyser la capacité de détection quantitative d' E." sur tous les triplets d'un groupe d'entités déterminé. Pour plus de clarté, nous nous référerons à ce jeu de données par le terme **'groupe A'**.

Nous utiliserons également le banc d'essai de Stanovsky et Dagan [13] afin d'analyser la capacité qualitative (rappel et précision) de ces heuristiques. Nous nous référerons à ce jeu de données par le terme **'groupe B'**.

4.1.1 Description des heuristiques et analyse des résultats

Les heuristiques que nous voulons analyser dans le cadre de ce mémoire peuvent se classer en trois catégories : les heuristiques d'union et intersection des triplets de sortie de différents OIE, les heuristiques de sélection par score de confiance et les heuristiques de sélection par S.

Comme ces heuristiques sont nombreuses et nous avons différentes sources des triplets ('groupe A' et 'groupe B') auxquels nous appliquons différentes heuristiques, nous avons cru bon de représenter les étapes de la tâche de filtrage dans le schéma 4.5.

Afin d'avoir une idée de la précision et du rappel de nos heuristiques, nous ne pouvons pas utiliser les triplets des vingt entités d'exemple que nous avons utilisé jusqu'à maintenant. Pour ce faire, nous utiliserons le banc d'essai de Stanovsky et Dagan [13], bien que celui-ci ne soit pas représentatif de la composition typique d'une KB.

Comme nous l'avons déjà mentionné, le banc d'essai de Stanovsky et Dagan utilise le corpus QA-SRL de He et Lewis et Zettlemoyer [7]. Stanovsky et Dagan ont établi les 3200 phrases brutes de ce corpus comme leurs phrases d'entrée. Ces 3200 phrases ont été tirées de deux sources différentes : Newswire (PropBank) et Wikipédia.

Or toutes nos heuristiques de sélection par S se basent sur l'analyse directe ou indi-

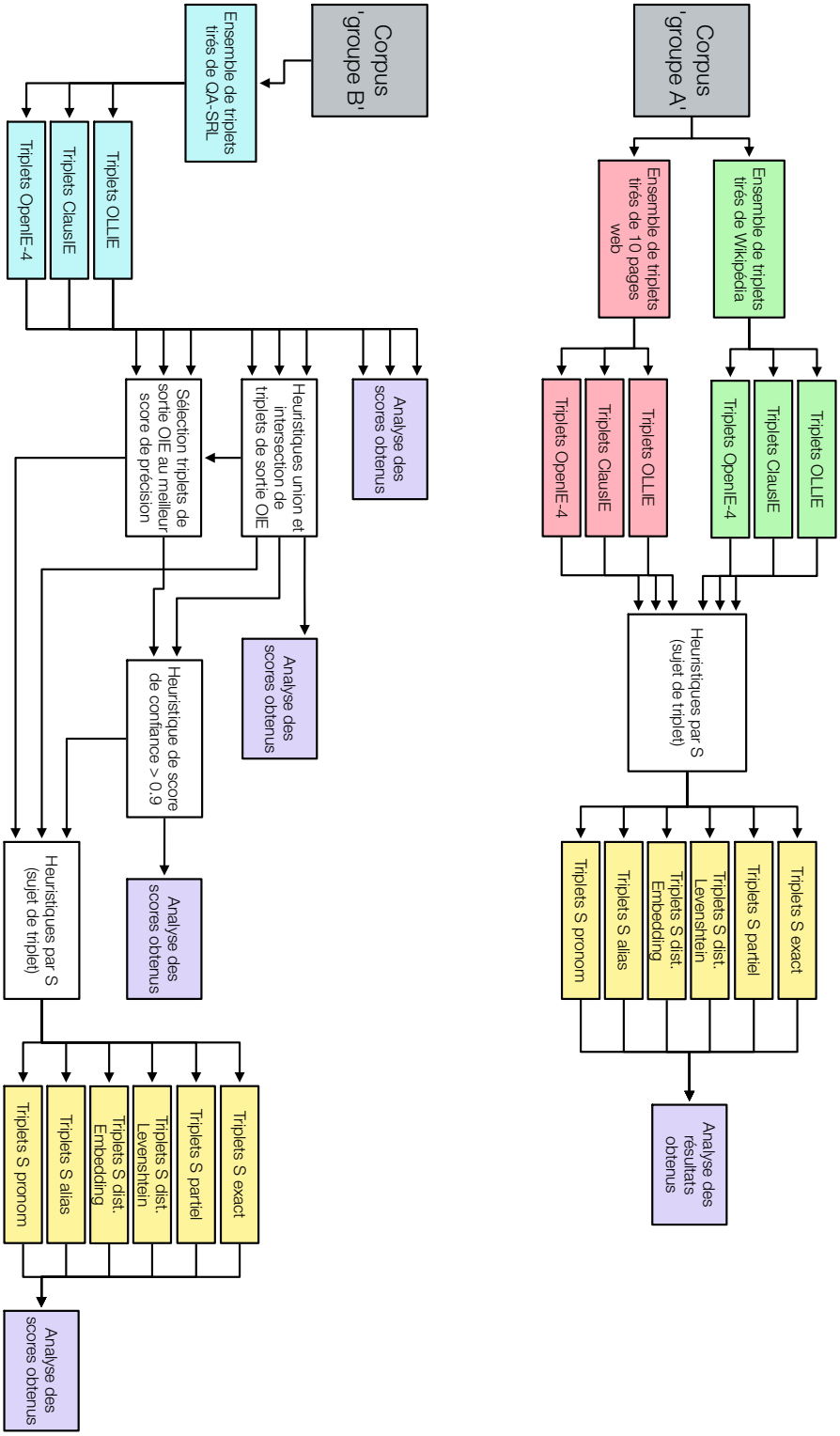


Figure 4.5 : Étapes suivies pour la tâche de filtrage sur les triplets de sortie OIE extraits du contenu Wikipédia et 10 pages web suggérées par Google Search (’groupe A’) ainsi que des phrases QA-SRL du banc d’essai (’groupe B’).

recte de l'entité E et du S de chaque triplet. Nous ne pouvons pas comparer le S de sortie OIE et l'entité E de chaque phrase brute du corpus, car E n'est pas précisée.

La solution que nous avons trouvée à ce problème est de n'utiliser que les 1959 phrases du corpus correspondant aux phrases extraites de Wikipédia et de faire, en amont, une recherche automatique par moteur de recherche. Pour cette recherche automatique nous utilisons les moteurs de recherche Bing (généreux en nombre d'autorisations d'usage) et Google Search (dans les cas où Bing est insuffisant) pour retrouver l'article Wikipédia d'où la phrase du corpus a été originalement extraite.

L'extraction des phrases du corpus QA-SRL date de 2006, c'est pourquoi dans certains cas les phrases originales ont été supprimées ou modifiées de l'article ou bien c'est l'article tout entier qui a été supprimé de Wikipédia. Pour cette raison, des 1959 phrases du corpus extraites de Wikipédia, nous n'avons gardé que 1927 d'entre elles. Ce sont ces 1927 phrases que nous avons établies comme notre corpus d'entrée pour le banc d'essai.

Afin d'effectuer une comparaison uniforme, nous avons utilisé ce corpus d'entrée pour analyser toutes nos heuristiques, qu'elles soient ou non basées sur la comparaison entre S et E. Et par souci de transparence et d'uniformité nous avons également reproduit la figure de courbes de précision et rappel des OIE de l'article de Stanovsky et Dagan [13] (cf. annexe I) en utilisant exclusivement les mentionnées 1927 phrases brutes pour corpus d'entrée (cf. la figure 4.6).

4.1.2 Description des heuristiques d'union et intersection de triplets extraits de différents OIE

Comme nous avons déjà expliqué, le jeu de triplets extraits par OIE a comme caractéristique d'être grand en nombre et abondant en bruit.

Bien que les OIE choisis utilisent différentes approches et algorithmes, l'objectif final de tous les OIE est de produire un même résultat : des triplets se rapprochant le plus possible de ce qu'un être humain pourrait juger sémantiquement pertinent. C'est d'ailleurs ce qui nous permet d'utiliser un banc d'essai pour juger de l'efficacité de

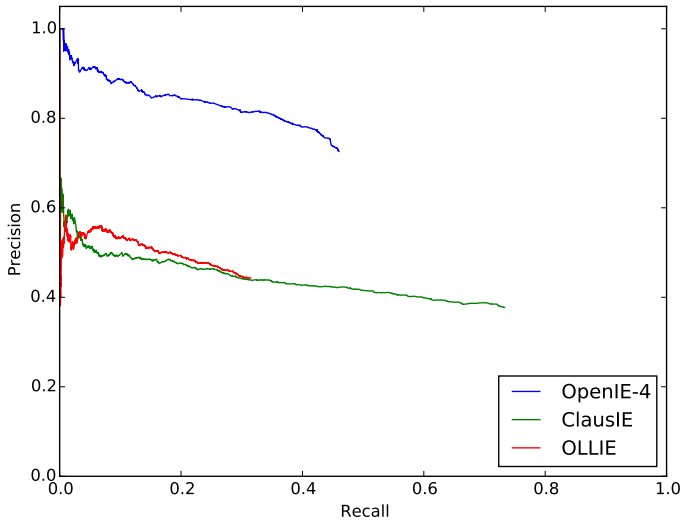


Figure 4.6 : (’groupe B’ au corpus d’entrée et au ’Gold Standard’ limité aux phrases extraites de Wikipédia) Figure de courbes mesurant la Précision et le Rappel en variant le seuil de confiance pour 3 OIE : OLLIE, ClausIE et OpenIE-4.

plusieurs OIE avec un seul ’Gold Standard’.

Le fait que chaque OIE utilise un ensemble de règles développé indépendamment et différemment des autres OIE pour extraire ses triplets, nous pousse à nous demander :

- est-ce que l’intersection des triplets de sortie de différents OIE permettrait de filtrer davantage de triplets non pertinents et capturer plus de triplets pertinents ?
- est-ce que l’union des triplets de sortie de différents OIE permettrait d’augmenter le nombre de triplets ayant un haut score de confiance (facilement filtrables par l’heuristique de score) ?

Il nous faut vérifier cette conjecture en analysant le score de précision obtenu sur le banc d’essai. Nous en obtenons les figures 4.7.

Nous remarquons sur ces figures que les intersections de triplets de deux OIE différents n’offre pas un score (au seuil de confiance le plus bas) remarquablement meilleur que le score obtenu par l’OIE OpenIE-4 tout seul (cf. tableau 4.I). Et que les unions de

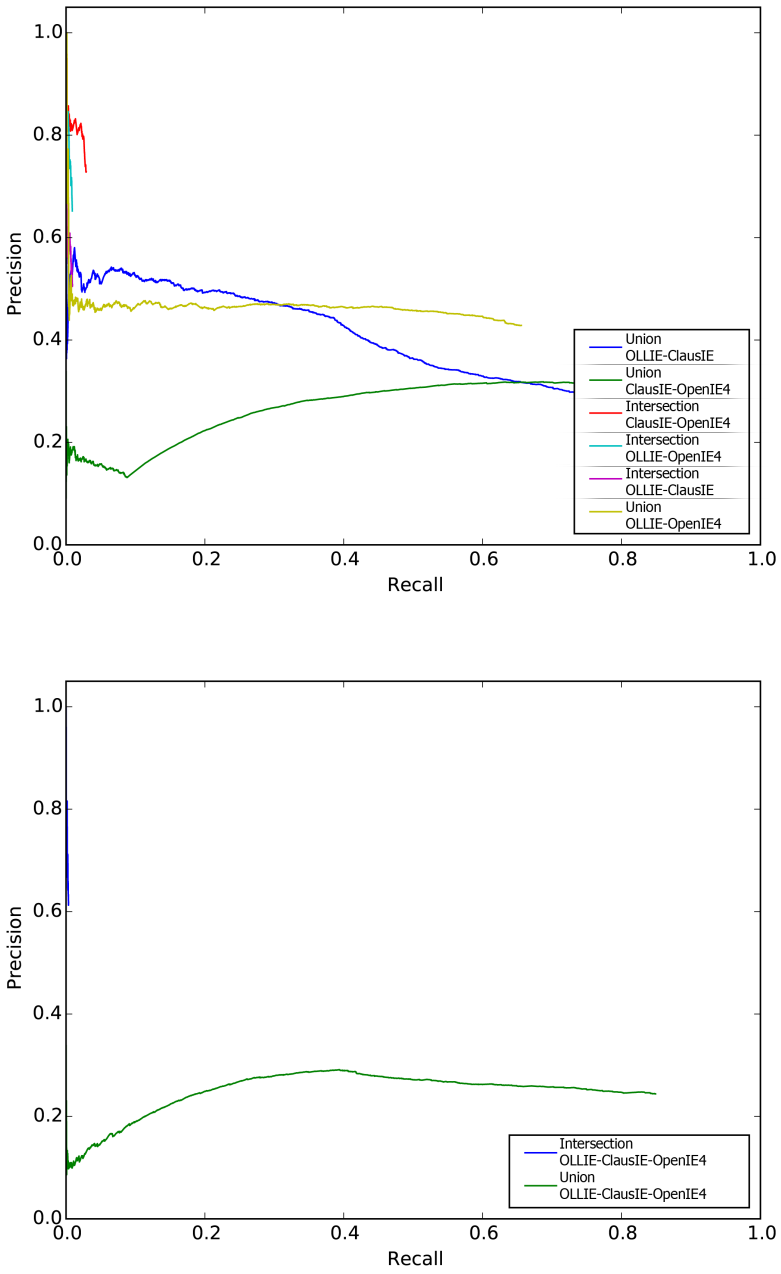


Figure 4.7 : (’groupe B’ au corpus d’entrée et au ’Gold Standard’ limité aux phrases extraites de Wikipédia) Figures de courbes mesurant la Précision et le Rappel en variant le seuil de confiance après l’application de diverses heuristiques (heuristiques d’union et intersection de triplets de sortie OIE).

triplets de deux OIE différents mélangent des triplets aux scores de confiance trop différemment déterminés pour être compatibles comme un seul jeu de données. C'est-à-dire que les scores de confiance (fournis par les OIE mêmes) ne sont pas équivalents [2].

	Precision
OLLIE	0.44
ClausIE	0.38
OpenIE-4	0.73
Intersection (OLLIE ∩ ClausIE)	0.51
Intersection (ClausIE ∩ OpenIE-4)	0.73
Intersection (OLLIE ∩ OpenIE-4)	0.65
Intersection (OLLIE ∩ ClausIE ∩ OpenIE-4)	0.62
Union (OLLIE ∪ ClausIE)	0.29
Union (ClausIE ∪ OpenIE-4)	0.32
Union (OLLIE ∪ OpenIE-4)	0.43
Union (OLLIE ∪ ClausIE ∪ OpenIE-4)	0.24

Tableau 4.I : ('groupe B' au corpus d'entrée et au 'Gold Standard' limité aux phrases extraites de Wikipédia) Tableau du score de précision au seuil de confiance le plus bas après l'application de diverses heuristiques (heuristiques d'union et d'intersection de triplets de sortie OIE).

Après l'analyse de ces figures et du tableau, nous remarquons que les meilleurs scores de précision sont ceux obtenus par l'OIE OpenIE-4 (isolément) et par l'intersection des triplets de sortie OIE ClausIE et OpenIE-4. La différence de score entre les deux est de l'ampleur de 0.25%, mais la différence de rappel est conséquente. Cette heuristique n'offre donc pas une grande amélioration de la précision et elle filtre de très nombreux triplets sur lesquels nous pourrions utiliser d'autres heuristiques.

[2]Dans ce cas, nous nous attendions à avoir une certaine équivalence entre les scores de confiance de OLLIE et de OpenIE-4 puisqu'ils ont été développés par le même groupe de recherche. Mais il semblerait que pendant le temps séparant le développement de OLLIE et OpenIE-4, la formule calculant la confiance de l'OIE ait changé et un score confiance de 0.9 chez OLLIE n'est pas aussi fiable qu'un score de 0.9 chez OpenIE-4.

Nous déterminons que ces heuristiques d'union et d'intersection des triplets de sortie OIE ne contribuent pas très notablement au filtrage heuristique pour un meilleur score de précision. Mais, par souci de méticulosité, nous testerons les heuristiques des scores de confiance et de S sur les triplets de sortie d'OpenIE-4 ainsi que sur les triplets de sortie de l'heuristique d'intersection de ClausIE et OpenIE-4.

4.1.3 Heuristiques par score de confiance

Si nous observons les figures produites par le banc d'essai de Stanovsky et Dagan [13], nous remarquerons que les courbes de précision et rappel sont généralement décroissantes. Ceci est dû au fait que le banc d'essai extrait les scores de confiance retournés par les OIE, les ordonne par ordre décroissant, les groupe par seuils de confiance et calcule pour chaque groupe la précision et le rappel avant de les transposer dans la figure. Nous remarquons que, souvent, il y a une corrélation entre le seuil de confiance et le score de précision. Bien sûr, les courbes montrent des fluctuations plus ou moins abruptes, mais la tendance générale est que plus le score de confiance est grand, plus il y a de probabilités pour que le triplet soit pertinent.

En nous servant de cette donnée, nous pouvons utiliser les scores pour nous défaire d'une grande partie des triplets non pertinents, bien que cette heuristique élimine aussi une quantité de triplets pertinents conséquente.

Après avoir appliqué cette heuristique, nous observons que les triplets de sortie OpenIE-4 ont un score de précision supérieur à 3% du score de l'intersection des triplets de sortie OpenIE-4 et ClausIE (cf. figure 4.8 et tableau 4.II).

Cette heuristique permet de filtrer très facilement les triplets et d'obtenir un meilleur score de précision. Malheureusement, ni cette heuristique ni la précédente ne permettent de filtrer les triplets ne ressemblant pas aux faits de KB. Il nous faut donc faire appel aux heuristiques par S. À nouveau, par souci de transparence, nous testerons les heuristiques suivantes (heuristiques par S) sur les triplets de sortie OIE sans appliquer l'heuristique de score et après l'avoir appliquée.

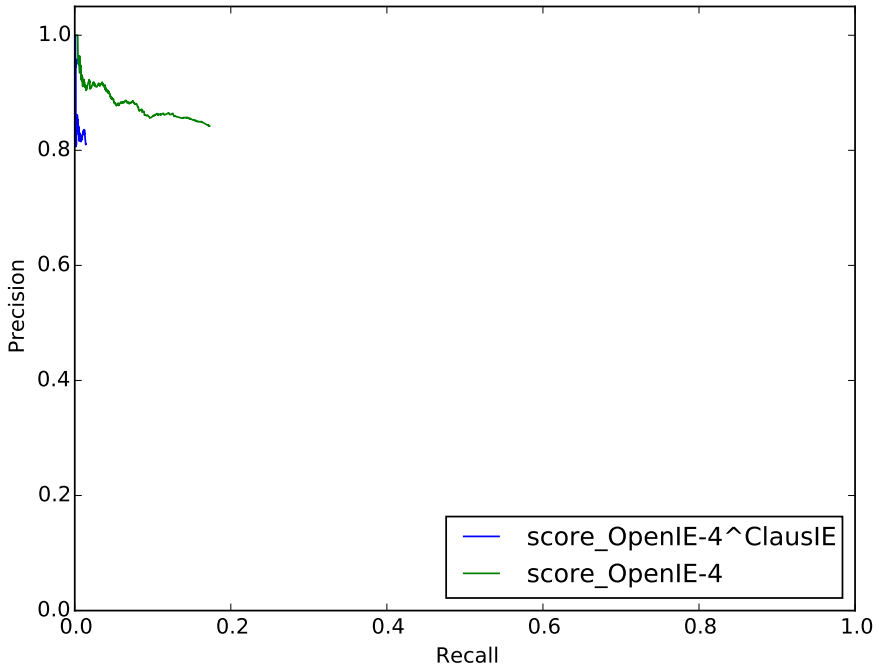


Figure 4.8 : (’groupe B’ au corpus d’entrée et au ’Gold Standard’ limité aux phrases extraites de Wikipédia) Figures de courbes mesurant la Précision et le Rappel avant et après l’application de diverses heuristiques (heuristique intersection OIE, heuristique score de confiance > 0.9).

	Precision
OpenIE-4	0.84
Intersection (ClausIE ∩ OpenIE-4)	0.81

Tableau 4.II : (’groupe B’ au corpus d’entrée et au ’Gold Standard’ limité aux phrases extraites de Wikipédia) Tableau du score de précision au seuil de confiance le plus bas après l’application d’une heuristique score de confiance > 0.9 (ne capturant que les triplets ayant un score de confiance supérieur à 0.9).

4.1.4 Heuristiques par sujet de triplet

Ces heuristiques sont nécessaires à l’heure d’assimiler les triplets de sortie OIE à des faits de KB. Cependant, il est bon de remarquer que, contrairement aux heuristiques précédentes, elles demandent davantage de réglages et d’expérimentation pour pouvoir trouver la ou les heuristiques optimales. Ceci est dû au fait qu’il nous faut trouver le meilleur moyen pour faire coïncider la multitude de possibilités (E’) jugées acceptables et la multitude de possibilités apparaissant comme S de triplets de sortie OIE.

Comme nous l’avons déjà expliqué et constaté dans les tableaux de l’annexe VII, les P des triplets varient grandement d’une KB à l’autre. De la même façon, nous avons observé par le tableau 3.I que les informations des faits diffèrent entre 93 et 95% d’une KB à l’autre, que se soit en forme ou en contenu. Cependant, nous pouvons remarquer que, par la définition même des KB, les faits d’une requête E auront invariablement un S identique à E.

Ce n’est pas le cas des sorties de triplets OIE, car l’objectif des OIE est de produire des triplets sémantiquement correspondants au texte d’entrée, pas de discriminer ses triplets. Donc, si nous voulons capturer les triplets de sortie ayant un S se référant directement à E, il nous faut réaliser un filtrage des S.

Nous pourrions nous limiter à ne garder que les triplets de sortie OIE dont le S correspond à E mot pour mot, caractère à caractère. Mais dans les textes libres, il existe généralement plusieurs façons de se référer à E et ces variantes synonymiques d’ E. sont

nombreuses et très communes. Par exemple, il est possible que l'entité du 'groupe A' "Chilly Gonzales" soit référée comme "Gonzales", "Jason Charles Beck", "Jason Beck", "Beck", "Gonzo", "Chilly", "Chilly G", "Chilly Gonzo", etcætera.

Si nous analysons exclusivement les triplets de sortie OIE ayant E pour S, nous obtenons un nombre très réduit de triplets jugés pertinents, car nous ne prenons pas en compte les coréférents (E') de l'E. (cf. les tableaux 4.III et 4.IV).

Heuristique	OLLIE	CLausIE	OpenIE-4
S exact (S == E)	28	10	7
TOTAL	6719	8134	2804

Tableau 4.III : ('groupe A') Tableau du nombre de triplets de sortie OIE ayant E pour S exact, obtenu à partir du corpus des dix premières suggestions de Google Search.

Heuristique	OLLIE	CLausIE	OpenIE-4
S exact (S == E)	30	12	6
TOTAL	1262	1120	445

Tableau 4.IV : ('groupe A') Tableau du nombre de triplets de sortie OIE ayant E pour S exact, obtenu à partir du corpus de l'article Wikipédia correspondant.

Ceci représente un problème bien connu dans les recherches en résolution de coréférence, mais qui prend ici un niveau de difficulté supplémentaire notamment dans le cas des textes libres choisis en utilisant Google Search dû au manque de transition et de constance entre les textes de différentes sources.

Il existe plusieurs outils permettant, jusqu'à un certain point, de résoudre la coréférence. Cependant leur fiabilité laisse parfois à désirer et leur utilisation risque d'introduire davantage de bruit dans les triplets de sortie OIE.

C'est pour cette raison que nous utilisons des heuristiques simples pour reconnaître les S correspondant à E' dans les triplets de sortie OIE.

Il est impossible d'être certains à cent pour cent que toutes nos heuristiques capturent exclusivement des triplets correspondant à E. Mais en nous débarrassant des triplets dont

nous sommes certains qu'ils ne correspondent pas à E, la concentration de triplets de sortie OIE pertinents augmente considérablement.

Dit autrement, à l'heure de lister les triplets ayant E pour S, nous incluons également les S correspondant à E' (coréférents de E).

En analysant les différents types de S pouvant correspondre à E, il nous est possible de les classifier en quatre groupes (selon les différents types de correspondance entre S et E) : entité partielle, entité similaire, alias et pronom.

Nous entendons par **"entité partielle"** le S composé d'une partie de E. Par exemple, pour l'entité "Chilly Gonzales" nous pouvons mentionner "Chilly" ou "Gonzales". Afin de tenir compte de ce type de S nous pouvons utiliser chacun des mots composant E pour trouver E'.

Nous entendons par **"entité similaire"** un S composé d'une variante orthographique de E. Par exemple, pour l'entité "Chilly Gonzales" nous pouvons mentionner "Chili Gonzales" ou "Chilly Gonzalez". Comme nous dépendons d'une classification non prédéfinie pour détecter ce type de S, cette méthode est plus arbitraire et moins fiable.

Il est toujours possible d'obtenir un score de distance Levenshtein entre le S de sortie OIE et E ou utiliser des Word embeddings pour calculer le score de similarité entre le vecteur de S et le vecteur de E. La difficulté réside dans le choix du score limite qui détermine la frontière entre ce qui est accepté et ce qui est rejeté. À partir de nos observations et de tests empiriques sur des centaines de calculs de distance, nous avons estimé que le score frontière optimal tourne dans les alentours de 70 et 80% que ce soit pour le score de distance Levenshtein ou le score de distance par Word embeddings.

Comme il nous faut choisir un score spécifique comme score frontière, nous avons choisi de l'établir à 75%. Cela ne veut pas dire que les couples E-S aux mesures de similarités ayant obtenu un score supérieur au score frontière ne contiennent aucune erreur ou que tous ceux en dessous du score frontière ne contiennent pas de couples réellement similaires, mais mal classifiés. Mais cela nous pousse à croire que la grande majorité des couples E-S du triplet sémantiquement similaires auront un score de similarité supé-

rieur au score frontière et que la grande majorité des couples sémantiquement différents auront un score de similarité inférieur au score frontière.

Nous entendons par **"alias"** un S composé d'un nom ou pseudonyme ayant le même référent que E. Par exemple, pour l'entité "Chilly Gonzales" nous pouvons citer "Gonzo" ou "Jason Beck". Pour pouvoir recenser les alias d' E., nous utilisons les faits disponibles dans les KB. En cherchant dans les différentes KB si E possède des alias connus et les utiliser pour reconnaître E'.

Nous entendons par **"pronom"** tout mot grammatical préétabli appartenant à un groupe langagier pour servir de substitut à un entité nommée (E). Par exemple, pour l'entité "Chilly Gonzales" nous pouvons citer "he" ou "him". La détection du coréférent d'un pronom par un outil de résolution de coréférence dépend grandement de la syntaxe de la phrase et du contexte explicite ou déductible, ce qui explique que ces outils aient encore des difficultés à faire correspondre E' à E lorsque E' est un pronom.

4.2 Analyse des résultats des heuristiques

Le corpus de 1927 phrases brutes que nous avons utilisées pour le banc d'essai est composé de phrases extraites de centaines d'articles Wikipédia différents et pour chaque article (donc pour chaque E), le nombre de phrases extraites dépasse rarement les 6 phrases.

Bien qu'idéal pour d'autres projets, ce corpus n'est pas vraiment représentatif des triplets de sortie OIE typique de notre projet. Il nous faut reconnaître qu'il fonctionne très bien pour ce qui est d'évaluer les heuristiques d'union et intersection des triplets de sortie de différentes OIE et de l'heuristique du score de confiance, mais pour ce qui est d'évaluer les heuristiques par S, il ne s'adapte pas très bien. Pour mieux s'adapter, il aurait fallu que le corpus soit composé de centaines de phrases tirées d'un même article Wikipédia au lieu de cinq ou six.

Mais puisque c'est ce corpus d'entrée que Stanovsky et Dagan [13] ont choisi pour

leur banc d'essai, il nous est impossible d'en choisir un autre.

4.2.1 Analyses sur le 'groupe A'

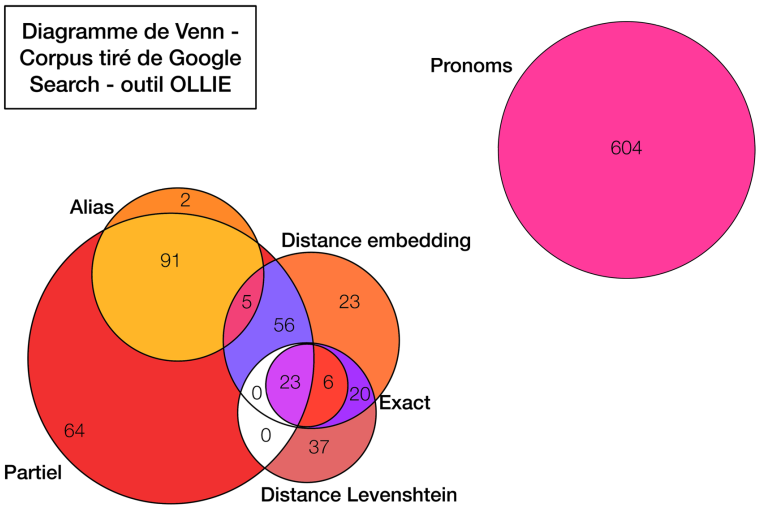
C'est pourquoi, afin d'avoir une meilleure idée du bon ou mauvais fonctionnement des heuristiques par S en matière de quantité, nous analyserons les résultats des heuristiques par S sur le 'groupe A'.

Nous observons sur les tableaux 4.V et 4.VI que le nombre de triplets que ces heuristiques sont capables de capturer est encore assez réduit en comparaison au nombre total de triplets, mais il est remarquablement plus important que lorsque nous nous limitons à capturer E et pas E'. Toutefois, nous pouvons nous demander si certaines heuristiques capturent les mêmes triplets que d'autres heuristiques ou s'il s'agit de triplets capturés par une seule heuristique et négligés par les autres.

Heuristique	OLLIE	CLausIE	OpenIE-4
S exact (S == E)	29	10	8
S partiel	182	216	92
S proche distance Levenshtein	64	43	22
S proche distance Embedding	133	126	58
S alias de E	96	87	51
S pronom	604	800	265
Toutes heuristiques par S confondues	852	1066	388
TOTAL	6719	8134	2804

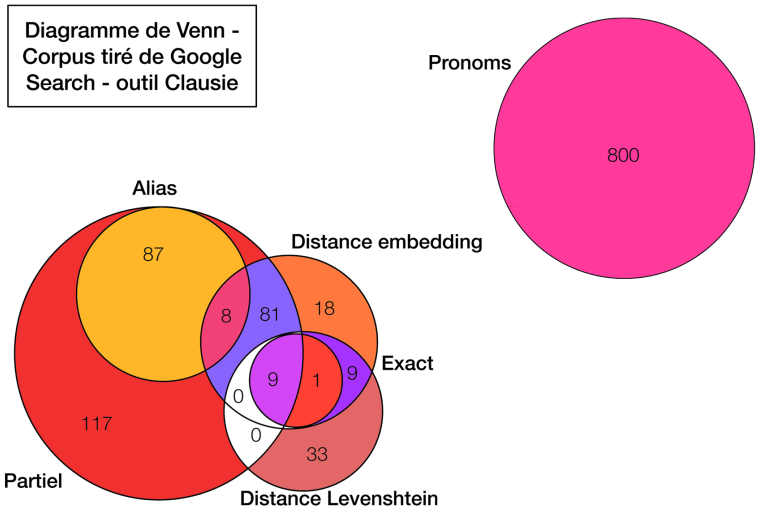
Tableau 4.V : ('groupe A') Tableau du nombre de triplets de sortie OIE après l'application de diverses heuristiques par S, à partir du corpus obtenu par Google Search.

Afin de vérifier quels triplets sont capturés par quelles heuristiques et vérifier si plusieurs heuristiques ont capturé les mêmes triplets, nous avons fait une analyse approfondie que nous avons représentée dans les figures 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 sous forme de diagrammes de Venn.



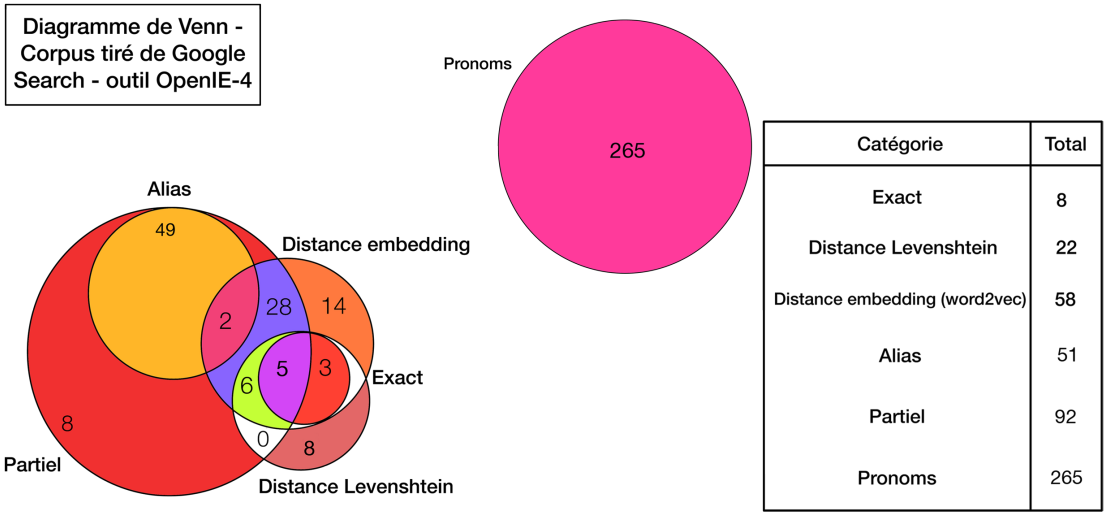
Catégorie	Total
Exact	29
Distance Levenshtein	64
Distance embedding (word2vec)	133
Alias	96
Partiel	182
Pronoms	604

Figure 4.9 : (’groupe A’) Diagrammes de Venn représentant le nombre de triplets capturés par différentes heuristiques pour les triplets de sortie OLLIE à partir du corpus obtenu par Google Search (diagrammes à l’aire non proportionnelle au nombre représenté).



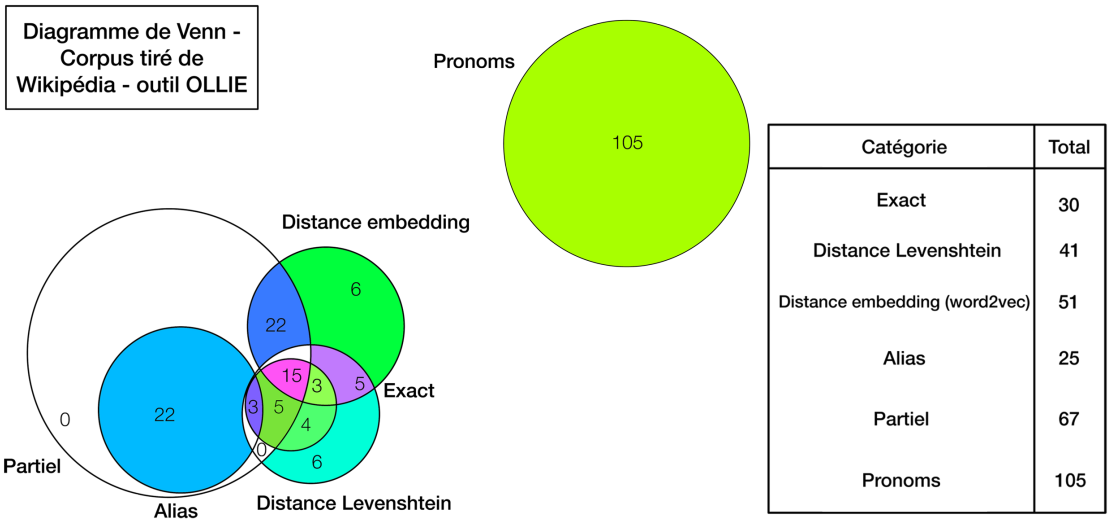
Catégorie	Total
Exact	10
Distance Levenshtein	43
Distance embedding (word2vec)	126
Alias	87
Partiel	216
Pronoms	800

Figure 4.10 : (’groupe A’) Diagrammes de Venn représentant le nombre de triplets capturés par différentes heuristiques pour les triplets de sortie ClausIE à partir du corpus obtenu par Google Search (diagrammes à l’aire non proportionnelle au nombre représenté).



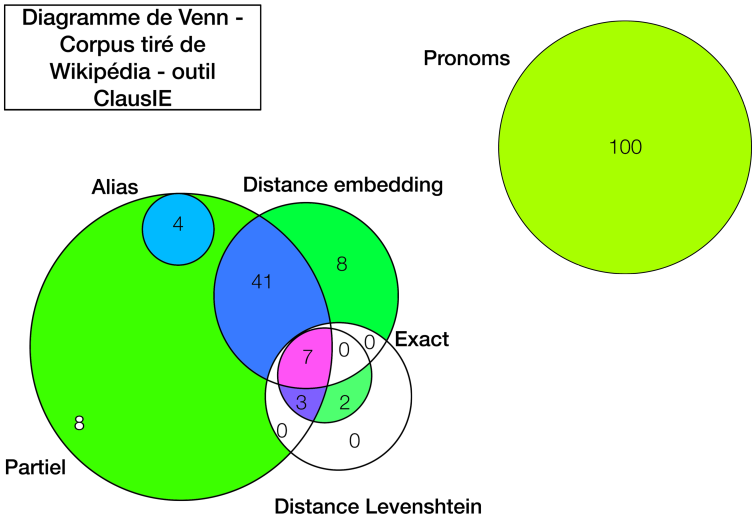
Catégorie	Total
Exact	8
Distance Levenshtein	22
Distance embedding (word2vec)	58
Alias	51
Partiel	92
Pronoms	265

Figure 4.11 : (’groupe A’) Diagrammes de Venn représentant le nombre de triplets capturés par différentes heuristiques pour les triplets de sortie OpenIE-4 à partir du corpus obtenu par Google Search (diagrammes à l’aire non proportionnelle au nombre représenté).



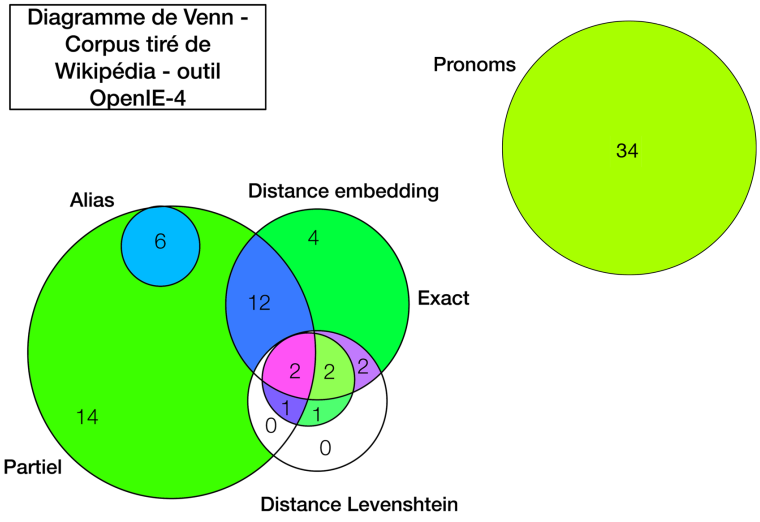
Catégorie	Total
Exact	30
Distance Levenshtein	41
Distance embedding (word2vec)	51
Alias	25
Partiel	67
Pronoms	105

Figure 4.12 : (’groupe A’) Diagrammes de Venn représentant le nombre de triplets capturés par différentes heuristiques pour les triplets de sortie OLLIE à partir de l’article Wikipédia (diagrammes à l’aire non proportionnelle au nombre représenté).



Catégorie	Total
Exact	12
Distance Levenshtein	17
Distance embedding (word2vec)	56
Alias	4
Partiel	63
Pronoms	100

Figure 4.13 : (’groupe A’) Diagrammes de Venn représentant le nombre de triplets capturés par différentes heuristiques pour les triplets de sortie ClausIE à partir de l’article Wikipédia (diagrammes à l’aire non proportionnelle au nombre représenté).



Catégorie	Total
Exact	6
Distance Levenshtein	8
Distance embedding (word2vec)	22
Alias	6
Partiel	23
Pronoms	34

Figure 4.14 : (’groupe A’) Diagrammes de Venn représentant le nombre de triplets capturés par différentes heuristiques pour les triplets de sortie OpenIE-4 à partir de l’article Wikipédia (diagrammes à l’aire non proportionnelle au nombre représenté).

Heuristique	OLLIE	CLausIE	OpenIE-4
S exact (S == E)	30	12	6
S partiel	67	63	23
S proche distance Levenshtein	41	17	8
S proche distance Embedding	51	56	22
S alias de E	25	4	6
S pronom	105	100	34
Toutes heuristiques par S confondues	196	173	74
TOTAL	1262	1120	445

Tableau 4.VI : (’groupe A’) Tableau du nombre de triplets de sortie OIE après l’application de diverses heuristiques par S, obtenu à partir de l’article Wikipédia.

En analysant ces figures, nous pouvons faire quelques observations : dans tous les cas analysés, l’heuristique capturant les "S pronoms" ne capture aucun triplet également capturé par les autres heuristiques. Ceci est logique puisque les entités recherchées ont rarement une section, un alias ou un "S hautement similaire" coïncidant avec un pronom [3].

Nous remarquons aussi que les triplets au S correspondant point par point avec E sont toujours capturés par les heuristiques de similarité par distance Levenshtein et par distance embedding. Finalement, remarquons que les triplets capturés par l’heuristique de S alias correspondent vastement avec ceux capturés par l’heuristique de S partiels de E.

Après ces analyses, nous pouvons conclure que le corpus d’entrée d’OIE obtenu par Google Search est plus profitable que celui obtenu par l’article Wikipédia et ce pour plusieurs raisons :

- Le fait de ne pas avoir à se limiter à des E et au contenu d’articles Wikipédia

[3] Il existe, évidemment, quelques exceptions à cette règle (non représentées dans les figures) parmi lesquelles à mode d’exemple, nous pouvons mentionner l’entité "Her" (film), l’entité "Him" (pièce de théâtre), l’entité "It" (roman), etcætera.

permet de pouvoir faire une requête sur, virtuellement, n'importe quelle entité, qu'elles soient ou pas dans Wikipédia [4].

- Il est possible d'aller chercher autant de pages web que rendues par le moteur de recherche, ce qui fait autant de contenu duquel il est possible d'extraire des triplets. Cela nous assure la capture de triplets de davantage plus nombreux et variés. Ce qui est une bien meilleure représentation de la large variété d'opinions et pensées humaines.

- Cela nous permet d'exploiter de nouvelles sources d'information peu utilisées par la plupart des autres KB.

Bien sûr, en effectuant ce choix nous sommes conscients qu'il n'est pas sans défauts. Le corpus d'entrée d'OIE obtenu par l'article Wikipédia a aussi des points forts qui font défaut à celui obtenu par Google Search. Pour être bien conscients des limitations de ce choix pour de futures recherches, nous faisons mention de quelques-uns des défauts principaux :

- Cette méthode est fortement dépendante du moteur de recherche utilisé (quel qu'il soit) et de sa capacité à obtenir des pages web pertinentes vis-à-vis de l'E de requête.

- En acceptant des informations plus libres que celles du modèle Wikipédia, le contrôle sur le contenu est moindre et il est bien plus probable que l'information soit hors sujet, contradictoire, bruitée. Ce qui fera que les extractions soient elles aussi hors sujet contradictoires et bruitées.

[4] Nous rapellons que pour nos exemples et la sélection des E de notre échantillon nous nous sommes limités à des E coïncidant avec des titres d'articles Wikipédia, mais il est peu probable que toutes les requêtes d'un utilisateur réel correspondent à des titres d'articles Wikipédia.

4.2.2 Analyses sur le 'groupe B'

Pour ce qui est de l'évaluation qualitative des triplets capturés par nos heuristiques et obtenir un score de précision et rappel, il nous faut faire utiliser 'groupe B', bien que celui-ci ne s'adapte pas parfaitement aux triplets de sortie OIE typiques de notre projet.

Après être passés par le banc d'essai de Stanovsky et Dagan [13], nous analysons les scores de précision obtenus (cf. tableaux 4.VII et 4.VIII [5]) nous observons que le meilleur score de précision est obtenu sur les triplets de sortie de l'intersection ClausIE et OpenIE-4, avec l'heuristique par S de proche distance embedding (avec et sans l'heuristique de score de confiance supérieur à 0.9). Ce score est de 1.0, mais peut n'être que le résultat du hasard, car nous ne retrouvons pas cette tendance de plus haute précision dans d'autres résultats.

Le second score de précision le plus important est de 0.92 et correspond aux résultats obtenus sur les triplets de sortie de l'intersection ClausIE et OpenIE-4, avec l'heuristique par S alias de E. Contrairement à l'heuristique par S de proche distance embedding, le score de précision de cette heuristique reste toujours le premier ou deuxième meilleur score. Ce qui semble nous indiquer que l'heuristique par S alias de E est, parmi les heuristiques que nous proposons, la plus invariable en ce qui concerne l'obtention du meilleur score de précision.

Afin d'avoir une idée plus précise des triplets que nous capturons, nous pouvons observer sous forme de tableau les triplets de sortie de nos heuristiques rendant les meilleurs scores de précision (cf. tableaux 4.IX, 4.X et 4.XI).

Bien que peu nombreux par rapport au nombre total de triplets, nous observons que les triplets extraits par l'intersection de OpenIE-4 et ClausIE et capturés par nos heuristiques ont un haut degré de validité sémantique.

Sans nous laisser influencer par triplets des tableaux ci-dessus, dont les S corres-

[5] À cause du score de rappel très bas des triplets capturés par nos heuristiques de S, les courbes produites par le banc d'essai sont difficilement distinguables, c'est pourquoi nous avons transcrit les résultats sous forme de tableau.

Triplets d'entrée	Heuristique	Score précision	Score Rappel	Score F-mesure
OpenIE-4	S exact (S == E)	0.68	0.002	0.004
OpenIE-4	S partiel	0.74	0.01	0.02
OpenIE-4	S proche distance Levenshtein	0.72	0.01	0.01
OpenIE-4	S proche distance Embedding	0.76	0.0006	0.001
OpenIE-4	S alias de E	0.84	0.004	0.01
OpenIE-4	S pronom	0.56	0.01	0.01
OpenIE-4	S exact ∩ Score de confiance > 0.9	0.78	0.002	0.003
OpenIE-4	S partiel ∩ Score de confiance > 0.9	0.80	0.006	0.01
OpenIE-4	S proche distance Levenshtein ∩ Score de confiance > 0.9	0.79	0.004	0.01
OpenIE-4	S proche distance Embedding ∩ Score de confiance > 0.9	0.77	0.001	0.001
OpenIE-4	S alias de E ∩ Score de confiance > 0.9	0.88	0.003	0.01
OpenIE-4	S pronom ∩ Score de confiance > 0.9	0.0	0.0	0.0
OpenIE-4	Toutes heuristiques par S	0.6672078	0.0253048	0.05
OpenIE-4	Toutes heuristiques par S sauf S pronom	0.75	0.01	0.03
OpenIE-4	Toutes heuristiques par S ∩ Score de confiance > 0.9	0.81	0.01	0.02
OpenIE-4	Toutes heuristiques par S sauf S pronom ∩ Score de confiance > 0.9	0.81	0.01	0.02

Tableau 4.VII : ('groupe B' au corpus d'entrée et au 'Gold Standard' limité aux phrases extraites de Wikipédia) Tableau du score de précision, rappel et f-mesure au seuil de confiance le plus bas après l'application de diverses heuristiques (heuristique score de confiance > 0.9, heuristiques par S).

pondent avec leur E, à ce stade il y a encore un risque de trouver dans les triplets de sortie des heuristiques des S qui ne correspondent pas à E. Tout ce que nous pouvons espérer d'obtenir à la sortie des heuristiques c'est des triplets dont la probabilité de correspondance entre S et E est bien supérieure à celle des triplets de sortie OIE ; et les

scores de précision nous indiquent juste le niveau de validité sémantique que les différents triplets de sortie des heuristiques ont.

Toutes les heuristiques n'ont pas la même probabilité de capturer des S correspondant ou pas à E. Il est à parier l'heuristique de S exact capture des triplets au S parfaitement correspondant à E tandis que le l'heuristique au S pronom n'en capture que très peu. Mais comme nous pouvons le constater sur les tableaux 4.VII et 4.VIII, une haute ou basse correspondance avec E n'implique pas un haut score de précision. Cependant, nous sommes plutôt ravis de voir que l'heuristique de S alias de E (une heuristique à haute probabilité de correspondance entre S et E) obtient invariablement des bons scores de précision.

En attendant de réaliser (dans un travail futur) une annotation manuelle de correspondance entre les S des sujets de sortie des heuristiques et E, ceci fait de l'heuristique alias un candidat parfait pour l'assimilation entre les triplets de sortie OIE et les faits de KB.

Triplets d'entrée	Heuristique	Score précision	Score Rappel	Score F-mesure
ClausIE ∩ OpenIE-4	S exact (S == E)	0.55	0.0003	0.0006
ClausIE ∩ OpenIE-4	S partiel	0.75	0.001	0.002
ClausIE ∩ OpenIE-4	S proche distance Levenshtein	0.76	0.001	0.002
ClausIE ∩ OpenIE-4	S proche distance Embedding	1.0	0.0001	0.0002
ClausIE ∩ OpenIE-4	S alias de E	0.93	0.001	0.001
ClausIE ∩ OpenIE-4	S pronom	0.46	0.002	0.003
ClausIE ∩ OpenIE-4	S exact ∩ Score de confiance > 0.9	0.56	0.0002	0.001
ClausIE ∩ OpenIE-4	S partiel ∩ Score de confiance > 0.9	0.82	0.001	0.001
ClausIE ∩ OpenIE-4	S proche distance Levenshtein ∩ Score de confiance > 0.9	0.75	0.001	0.001
ClausIE ∩ OpenIE-4	S proche distance Embedding ∩ Score de confiance > 0.9	1.0	0.0001	0.0002
ClausIE ∩ OpenIE-4	S alias de E ∩ Score de confiance > 0.9	0.9	0.0004	0.001
OpenIE-4	S pronom ∩ Score de confiance > 0.9	0.0	0.0	0.0
ClausIE ∩ OpenIE-4	Toutes heuristiques par S	0.58	0.004	0.01
ClausIE ∩ OpenIE-4	Toutes heuristiques par S sauf S pronom	0.76	0.002	0.003
ClausIE ∩ OpenIE-4	Toutes heuristiques par S ∩ Score de confiance > 0.9	0.77	0.001	0.002
ClausIE ∩ OpenIE-4	Toutes heuristiques par S sauf S pronom ∩ Score de confiance > 0.9	0.77	0.001	0.002

Tableau 4.VIII : ('groupe B' au corpus d'entrée et au 'Gold Standard' limité aux phrases extraites de Wikipédia) Tableau du score de précision, rappel et f-mesure des triplets de sortie OpenIE-4 au seuil de confiance le plus bas après l'application de diverses heuristiques (heuristique intersection OpenIE-4 et ClausIE, heuristique score de confiance > 0.9, heuristiques par S).

E	S (E')	P	O	Phrase originale
KOMO-TV	KOMO-TV	to become	the first television station in the nation to broadcast in true color	His discovery allowed KOMO-TV to become the first television station in the nation to broadcast in true color .
Parragon	Parragon	publishes	over 2000 titles every year	Parragon publishes over 2000 titles every year in over 25 languages .
Reprise	Reprise	can refer	to a version of a song	Reprise can refer to a version of a song which is similar to , yet different from , the song on which it is based .

Tableau 4.IX : ('groupe B' au corpus d'entrée et au 'Gold Standard' limité aux phrases extraites de Wikipédia) Tableau montrant la E correspondant à chaque triplet de sortie de l'heuristique intersection OpenIE-4 et ClausIE et de l'heuristique par S proche distance embedding.

E	S (E')	P	O	Phrase originale
Cherokee language	Cherokee	has	many pronominal prefixes	Like many Native American languages , Cherokee has many pronominal prefixes .
Tyabb, Victoria	Tyabb	has	Tyabb Airport	Tyabb also has Tyabb Airport , a private airfield which has been operating for more than thirty years .
Prabhu Deva	Deva	has moved	to Mumbai	Deva has moved to Mumbai and is residing at Boney Kapoor 's old place called Green Acres .
Prabhu Deva	Deva	is residing	at Boney Kapoor 's old place	Deva has moved to Mumbai and is residing at Boney Kapoor 's old place called Green Acres .
John Keats	Keats	began	studying there in October 1815	Having finished his apprenticeship with Hammond , Keats registered as a medical student at Guy 's Hospital and began studying there in October 1815 .
Mark Hofmann	Hofmann	was born	in Salt Lake City	Hofmann was born in Salt Lake City , Utah .
Mark Hofmann	Hofmann	was	a below-average high school student	Hofmann was a below-average high school student , but he had many hobbies including magic , electronics , chemistry , and stamp and coin collecting .

Tableau 4.X : ('groupe B' au corpus d'entrée et au 'Gold Standard' limité aux phrases extraites de Wikipédia) Tableau 1/2 montrant la E correspondant à chaque triplet de sortie de l'heuristique intersection OpenIE-4 et ClausIE et de l'heuristique par S alias de E.

E	S (E')	P	O	Phrase originale
Noatak, Alaska	Noatak	has	a gravel public airstrip	Noatak has a gravel public airstrip and is primarily reached by air .
Gordie Howe	Gordie	could do	everything	When Richard retired in 1960 , he paid tribute to Howe , saying " Gordie could do everything . "
William Pitt, 1st Earl of Chatham	Pitt	spoke out	against the Convention of El Pardo	Pitt spoke out against the Convention of El Pardo which aimed to settle the dispute peacefully .
Scott McNeil	Scott	pursued	the theater	Even though he knew about voice acting , Scott pursued the theater .
Evel Knievel	Knievel	was scheduled	to jump a tank full of live sharks	Knievel was scheduled to jump a tank full of live sharks and would be televised live nationally .
Evel Knievel	Knievel	broke	his arms	Although Knievel broke his arms , he was more distraught over a permanent injury his accident caused to the cameraman .
James Hogue	James Arthur Hogue	is	a US impostor	James Arthur Hogue is a US impostor who most famously entered Princeton University by posing as a self-taught orphan .

Tableau 4.XI : ('groupe B' au corpus d'entrée et au 'Gold Standard' limité aux phrases extraites de Wikipédia) Tableau 2/2 montrant la E correspondant à chaque triplet de sortie de l'heuristique intersection OpenIE-4 et ClausIE et de l'heuristique par S alias de E.

CHAPITRE 5

CONCLUSION

Le travail réalisé pour le projet présenté dans ce mémoire nous a permis d'enrichir nos connaissances théoriques, de nous familiariser avec l'état de l'art en construction de KB et en développement d'OIE. Il nous a également permis d'acquérir de nouvelles compétences pratiques en extraction d'information et en programmation Python.

5.1 Description et résultats du projet

Pour mener à bien ce projet, nous avons progressé par plusieurs étapes que nous détaillons à la suite :

5.1.1 Observations

Afin de mieux comprendre notre objet d'étude, nous avons examiné la "composition" des KB et la forme et la quantité des triplets de sortie OIE. Nous avons observé des points communs et des points de divergence et nous avons spéculé les possibles manières d'assimiler les triplets de sortie OIE aux faits des KB.

5.1.2 Expériences

Nous avons réalisé plusieurs expériences correspondant à nos hypothèses afin de vérifier quelles heuristiques d'assimilation fonctionnent le mieux.

5.1.3 Les résultats obtenus

Les résultats rendus par nos systèmes de mesure indiquent que l'ensemble d'heuristiques rendant les meilleurs résultats sont :

- L'heuristique d'intersection des triplets de sortie OpenIE-4 et ClausIE ajoutée à l'heuristique par S proche distance embedding.

- L'heuristique d'intersection des triplets de sortie OpenIE-4 et ClausIE ajoutée à l'heuristique par S alias de E.

Le premier ensemble d'heuristiques rend un meilleur résultat ponctuel à deux reprises. Le second ensemble comprend l'heuristique par S alias de E. En théorie et selon nos observations empiriques, cette heuristique permet de capturer des triplets à S hautement correspondant à E. Qui plus est, tout au long des expériences, cette heuristique a montré une tendance à rendre le ou l'un des meilleurs résultats, indépendamment des autres heuristiques utilisées.

Pour un approfondissement dans les heuristiques, pour des reproductions partielles ou totales des expériences ou des recherches ultérieures, cette heuristique est à prendre en considération.

5.2 Perspectives

Par rapport aux objectifs que nous nous sommes tracés pour ce mémoire, nous jugeons avoir obtenu des résultats acceptables.

Cependant, nous sommes encore loin d'être complètement satisfaits, car il ne s'agit ici que d'une toute première approche (nécessaire) à l'assimilation de triplets de sortie OIE à des faits de KB.

Il existe plusieurs pistes que nous pouvons explorer dans des travaux futurs afin d'améliorer notre méthode et nos résultats. À la suite, nous vous donnons un début de liste de questions sur lesquelles nous aimerions continuer à travailler :

- Pouvons-nous approfondir davantage dans l'analyse statistique entre les faits des KB et les triplets de sortie OIE ?

- Devons-nous améliorer ou créer un banc d'essai nous permettant d'obtenir des résultats plus représentatifs de l'efficacité ou l'inefficacité des heuristiques proposées ?
- Pouvons-nous affiner ou trouver de meilleures heuristiques permettant de filtrer davantage les triplets de sortie OIE pour obtenir davantage de triplets sémantiquement corrects ?
- Pouvons-nous mettre en place des métaheuristiques fonctionnelles nous permettant de choisir parmi les heuristiques disponibles celles capturant le mieux et le plus de triplets selon le cas et l'origine de l'information libre ?
- Est-ce que des modèles statistiques entraînés sur les triplets corrects déjà capturés nous permettraient de mieux capturer de nouveaux triplets ?
- Est-ce que des réseaux de neurones entraînés sur les triplets corrects déjà capturés nous permettraient de mieux capturer de nouveaux triplets ?
- Existe-t-il un intérêt immédiat justifiant l'usage de ressources nécessaires à long terme pour mettre en place une KB contenant des faits éphémères, des appréciations, des états transitoires, des faits incertains, etcætera ?
- Pouvons-nous mettre en place un système permettant la classification des faits subjectifs de la KB décrite par branches d'opinion (époque, groupe culturel, région, personnalité), permettant de faire des associations entre les préférences appréciatives d'un utilisateur et telle ou telle branche des faits ?
- Est-ce qu'une KB contenant majoritairement des données subjectives complémenterait bien un système d'IA, un assistant personnel intelligent (comme Siri de Apple, Alexa de Amazon, Google de Google, Cortana de Microsoft) ou un système de question-réponse automatique (comme le Watson de IBM) pour tout ce qui a trait à la demande d'opinion sur un thème spécifique ?

BIBLIOGRAPHIE

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak et Sebastian Hellmann. Dbpedia-a crystallization point for the web of data. *Web Semantics : science, services and agents on the world wide web*, 7(3):154–165, 2009.

[2] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr et Tom M Mitchell. Toward an architecture for never-ending language learning. Dans *AAAI*, volume 5, page 3, 2010.

[3] Janara Christensen, Mausam, Stephen Soderland et Oren Etzioni. Semantic role labeling for open information extraction. Dans *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR '10, pages 52–60, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1866775.1866782`.

[4] Luciano Del Corro et Rainer Gemulla. Clausie : Clause-based open information extraction. Dans *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 355–366, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2035-1. URL `http://doi.acm.org/10.1145/2488388.2488420`.

[5] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun et Wei Zhang. Knowledge vault : A web-scale approach to probabilistic knowledge fusion. Dans *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM. ISBN

978-1-4503-2956-9. URL http://doi.acm.org/10.1145/2623330.2623623.

[6] Google. Freebase data dumps. URL https://developers.google.com/freebase/.

[7] Luheng He, Mike Lewis et Luke Zettlemoyer. Question-answer driven semantic role labeling : Using natural language to annotate natural language. Dans *EMNLP*, pages 643–653, 2015.

[8] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich et Gerhard Weikum. Yago2 : a spatially and temporally enhanced knowledge base from wikipedia. 2012.

[9] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland et Oren Etzioni. Open language learning for information extraction. Dans *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 523–534, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390948.2391009.

[10] Mausam Mausam. Open information extraction systems and downstream applications. Dans *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 4074–4077. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3061053.3061220.

[11] Tim O'REILLY. Freebase will prove addictive, 2007. URL http://radar.oreilly.com/2007/03/freebase-will-prove-addictive.html.

[12] Harinder Pal. Demonyms and compound relational nouns in nominal open ie. *Proceedings of AKBC*, pages 35–39, 2016.

[13] Gabriel Stanovsky et Ido Dagan. Creating a large benchmark for open information extraction. Dans *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, November 2016. Association for Computational Linguistics.

[14] Fabian M Suchanek, Gjergji Kasneci et Gerhard Weikum. Yago : A large ontology from wikipedia and wordnet. *Web Semantics : Science, Services and Agents on the World Wide Web*, 6(3):203–217, 2008.

[15] Denny Vrandečić. Wikidata : A new platform for collaborative data collection. Dans *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 1063–1064, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1230-1. URL `http://doi.acm.org/10.1145/2187980.2188242`.

[16] Fei Wu et Daniel S Weld. Open information extraction using wikipedia. Dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics, 2010.

[17] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead et Stephen Soderland. Textrunner : Open information extraction on the web. Dans *Proceedings of Human Language Technologies : The Annual Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, NAACL-Demonstrations '07, pages 25–26, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1614164.1614177`.

Annexe I

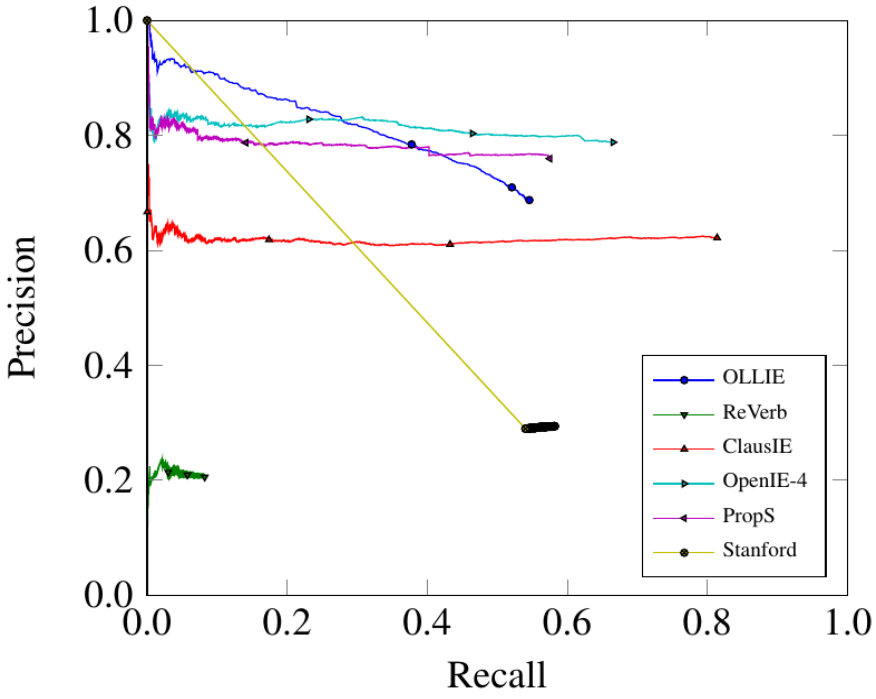


Figure I.1 : Schéma du banc d'essai extrait de Stanovsky et Dagan [13], courbe mesurant la Précision et le Rappel en variant le seuil de confiance pour 6 OIE : ReVerb, Stanford OpenIE, PropS, OLLIE, ClausIE et OpenIE-4

Annexe II

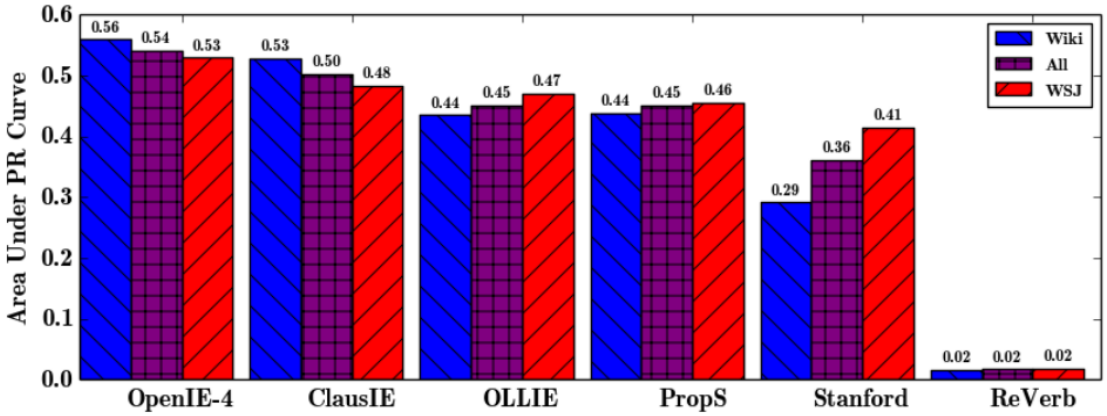


Figure II.1 : Schéma du banc d'essai extrait de Stanovsky et Dagan [13] analysant les scores de précision, rappel et aire sous la courbe pour plusieurs OIE

Annexe III

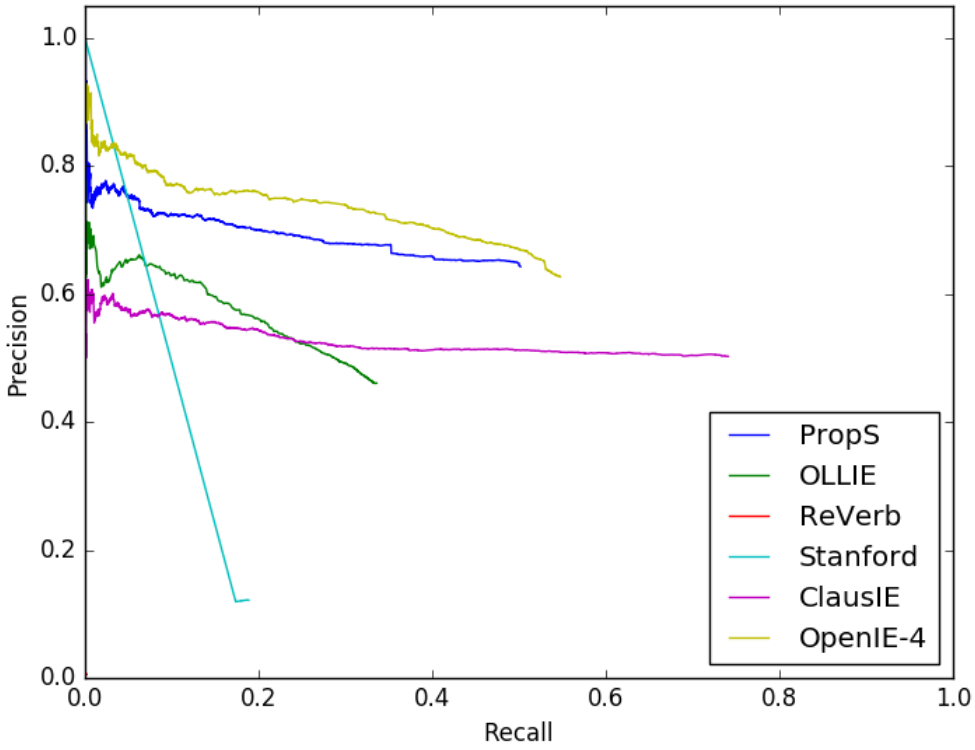


Figure III.1 : Schéma du banc d'essai extrait de https ://github.com/gabrielStanovsky/oie-benchmark, courbe mesurant la Précision et le Rappel en variant le seuil de confiance pour 6 OIE : ReVerb, Stanford OpenIE, PropS, OLLIE, ClausIE et OpenIE-4

Annexe IV

Wikidata se présente comme une collection d'entités[1] contenant des faits[2] pouvant contenir une information ou un lien vers une autre entité. Bien qu'il y ait certaines similarités entre ce format et le format de triplets RDF, Wikidata n'utilise pas l'ontologie RDF ni la composition sujet-propriété-objet (S-P-O) pour ses faits. Au lieu de cela, chaque entité a une seule instance qui contient tous les faits (P-O) correspondant à cette entité. Cette structure est fortement influencée par la structure traditionnelle du site Wikipédia et les objets informatiques de type dictionnaire.

[1] "Items" dans la terminologie Wikidata.

[2] "Statements" ou "affirmations" dans la terminologie Wikidata.

```
{mw.config.set({
    "wgPageName":"Q9309",
    {"wgUserGroups":["*"],
    "wbEntityId":"Q9309",
    "wbEntity":"{
        \"type\":\"item\",
        \"id\":\"Q9309\",
        \"labels\":{
            \"en\":{
                \"language\":\"en\",
                \"value\":\"Welsh\"},
                \"fr\":{
                 \"language\":\"fr\",
                    \"value\":\"gallois\"}
            }
        \"aliases\":{
            \"en\":[{
                \"language\":\"en\",
                \"value\":\"Welsh language\"Freebase
                },
                {
                \"language\":\"en\",
                \"value\":\"Cymric\"
                },
                {
                \"language\":\"en\",
                \"value\":\"cy\"
                }
            ]
        }
    }
    })
}
```


Figure IV.1 : Extrait de l'entité Wikidata Q9309 ("Welsh language") montrant les faits associés au type, au code d'identification, au nom principal et aux noms secondaires en plusieurs langues.

Annexe V

Les faits des sujets Freebase se présentent comme des entités liées à des ontologies et, dans leur forme, le Sujet et la Propriété correspondent à une autre entité de la KB tandis que l'Objet correspond soit à une information (chaîne de caractères), soit à un lien vers une entité.

```
<http://rdf.freebase.com/ns/m.083tk>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://rdf.freebase.com/ns/language.human_language>

<http://rdf.freebase.com/ns/m.083tk>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://rdf.freebase.com/ns/base.type_ontology.abstract>

<http://rdf.freebase.com/ns/m.083tk>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://rdf.freebase.com/ns/base.ontologies.ontology_instance>
de relations

<http://rdf.freebase.com/ns/m.083tk>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://rdf.freebase.com/ns/user.ktrueman.default_domain.official_language>

<http://rdf.freebase.com/ns/m.083tk>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://rdf.freebase.com/ns/base.welshculture.topic>

<http://rdf.freebase.com/ns/m.083tk>
    <http://rdf.freebase.com/ns/type.object.name>
        "Welsh Language"@en
```

Figure V.1 : Extrait de l'entité Freebase m.083tk ("Welsh language") montrant les faits associés aux types et nom en anglais.

Annexe VI

Chaque entité Knowledge Graph se présente (aux développeurs ayant accès à l'API Knowledge Graph) comme une collection d'entités rappelant le type d'objet informatique dictionnaire et contenant des faits (P-O) structurés.


```
{
    "@type": "EntitySearchResult",
    "result": {
        "@id": "kg:/g/120t40cc",
        "name": "Welsh Pony and Cob",
        "@type": [
            "Thing"
            ],
        "description": "Animal",
        "image": {
            "contentUrl": "http://t1.gstatic.com/images?q=tbn:
                ANd9GcSk5zNVktBcv45mdIehX_nMaN2t5KZLwo7GIy4fua0xIBF7y4zG",
            "url": "https://commons.wikimedia.org/wiki/File:WelshPonySectionD.jpg",
            "license": "http://creativecommons.org/licenses/by-sa/3.0"
            },
        "detailedDescription": {Duis Lacinia Rutrum
            "articleBody": "The Welsh Pony and Cob are a group of four closely
                related horse breeds including both pony and cob types, which
                originated in Wales in the United Kingdom. ",
            "url": "https://en.wikipedia.org/wiki/Welsh_Pony_and_Cob",
            "license": "https://en.wikipedia.org/wiki/Wikipedia:
                Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License"
            }
        },
    "resultScore": 59.911964
}
```


Figure VI.1 : Extrait exemplaire de l'entité Knowledge Graph g.120t40cc ("Welsh Pony and Cob") montrant les faits associés au code d'identification, au nom, à une courte description, à une description détaillée et à une image représentative.

Annexe VII

CORRESPONDANCE SUJET - PROPRIETE - OBJET

	Freebase	Knowledge Graph	Wikidata	TOTAL FAITS
Freebase	-	0	0	770
Knowledge Graph	-	-	2	250
Wikidata	-	-	-	394
TOTAL FAITS	770	250	394	

Tableau VII.I : Tableau de contingence montrant, pour les vingt entités d'échantillon, le nombre correspondances entre le sujet et la propriété et l'objet des faits contenus dans les trois KB choisies.

CORRESPONDANCE SUJET - PROPRIETE

	Freebase	Knowledge Graph	Wikidata	TOTAL FAITS
Freebase	-	0	0	770
Knowledge Graph	-	-	11	250
Wikidata	-	-	-	394
TOTAL FAITS	770	250	394	

Tableau VII.II : Tableau de contingence montrant, pour les vingt entités d'échantillon, le nombre correspondances entre le sujet et la propriété des faits contenus dans les trois KB choisies.

CORRESPONDANCE SUJET - OBJET

	Freebase	Knowledge Graph	Wikidata	TOTAL FAITS
Freebase	-	63	60	770
Knowledge Graph	-	-	39	250
Wikidata	-	-	-	394
TOTAL FAITS	770	250	394	

Tableau VII.III : Tableau de contingence montrant, pour les vingt entités d'échantillon, le nombre correspondances entre le sujet et l'objet des faits contenus dans les trois KB choisies.

Annexe VIII

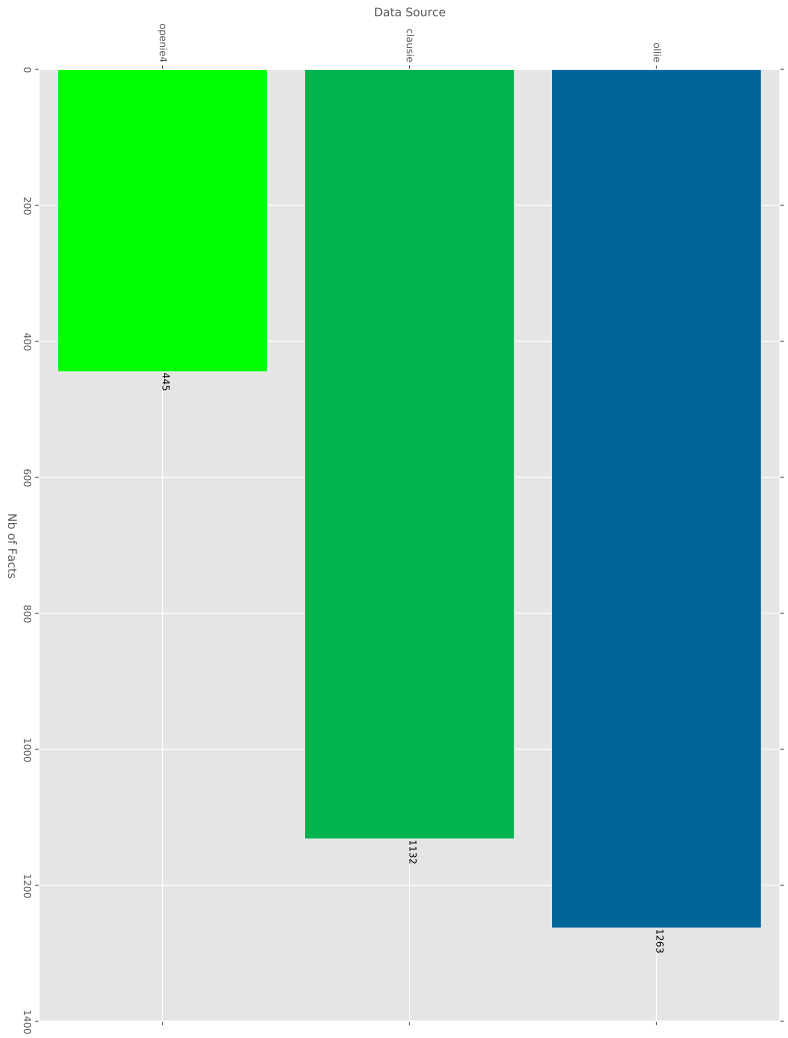


Figure VIII.1 : Nombre de triplets extraits du contenu de l'article Wikipédia, toutes entités confondues.

Annexe IX

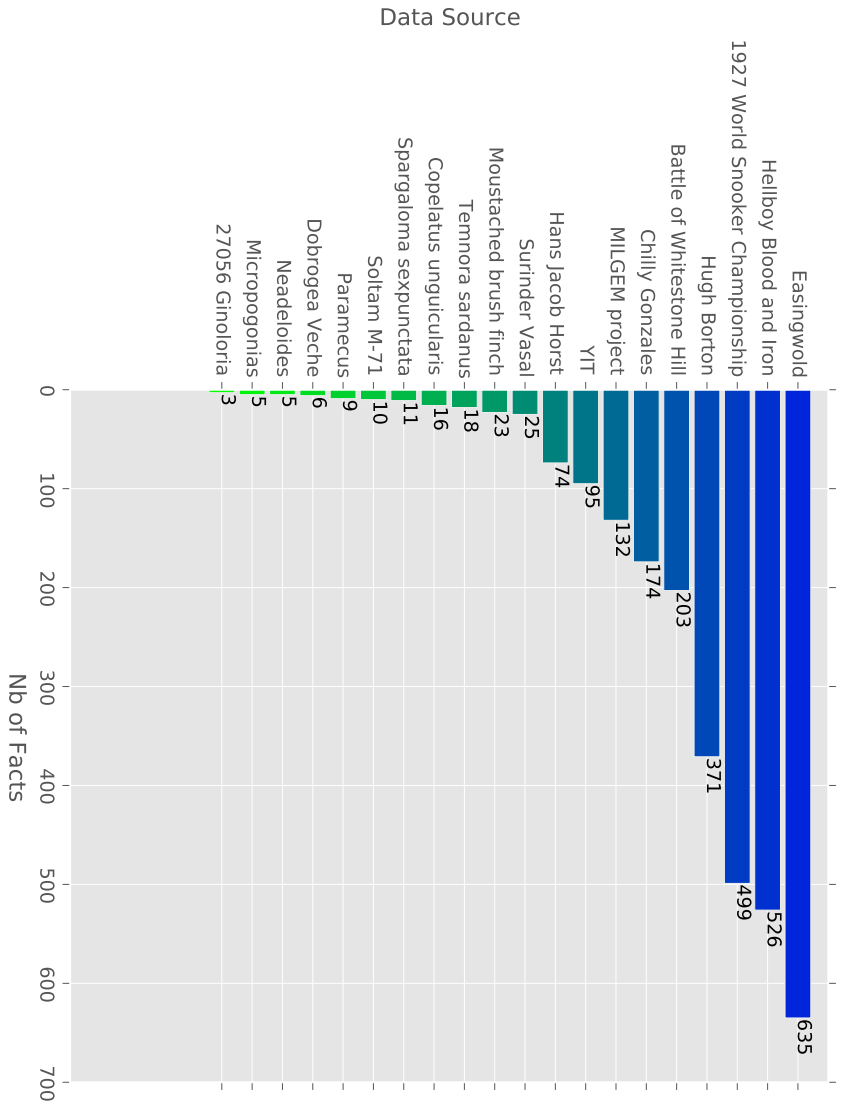


Figure IX.1 : Nombre de triplets extraits du contenu de l'article Wikipédia, segmenté par entités, toutes OIE confondues.

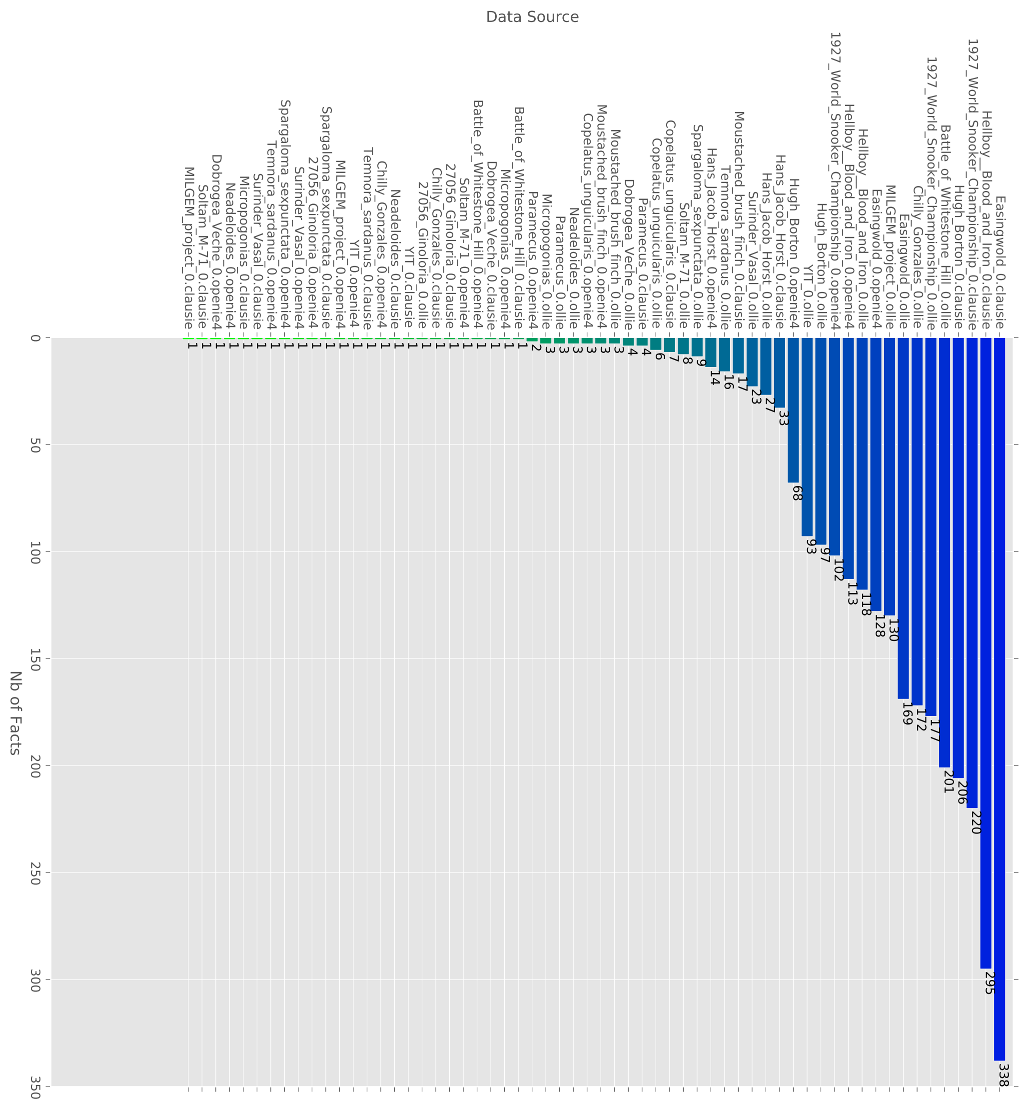


Figure X.1 : Nombre de triplets extraits du contenu de l'article Wikipédia, segmenté par OIE et par entité.

Annexe XI

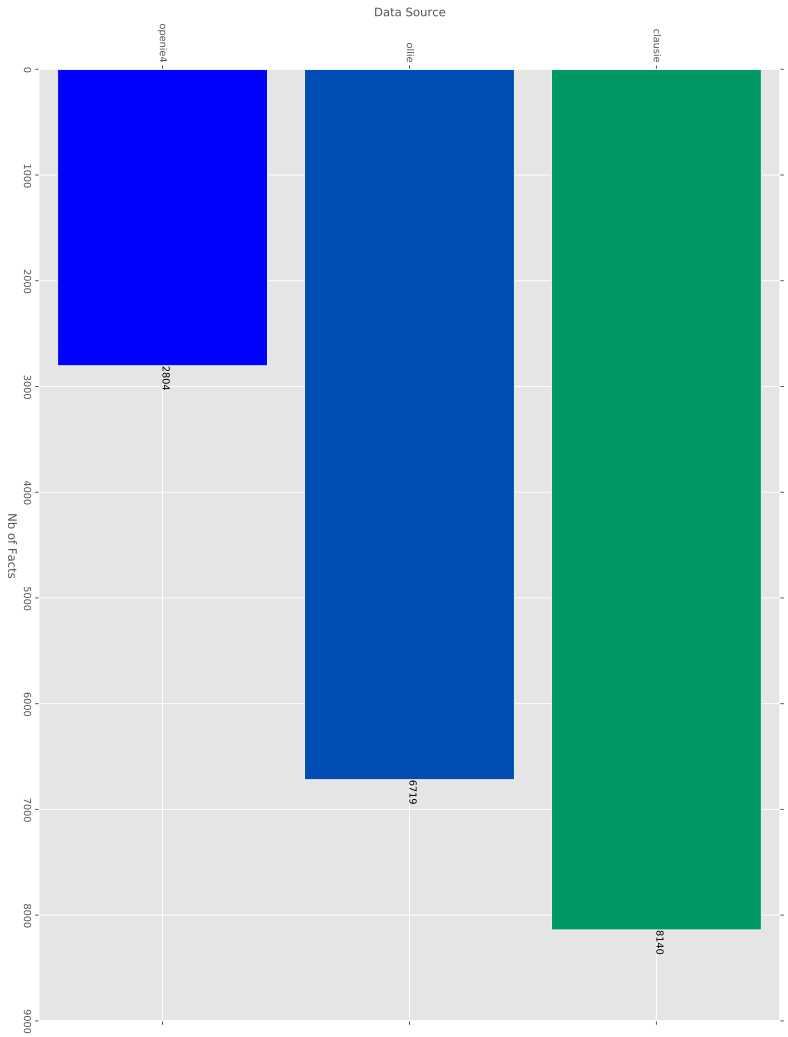


Figure XI.1 : Nombre de triplets extraits du contenu des 10 premières pages rendues par Google Search. Les triplets sont segmentés par OIE, toutes entités confondues.

Annexe XII

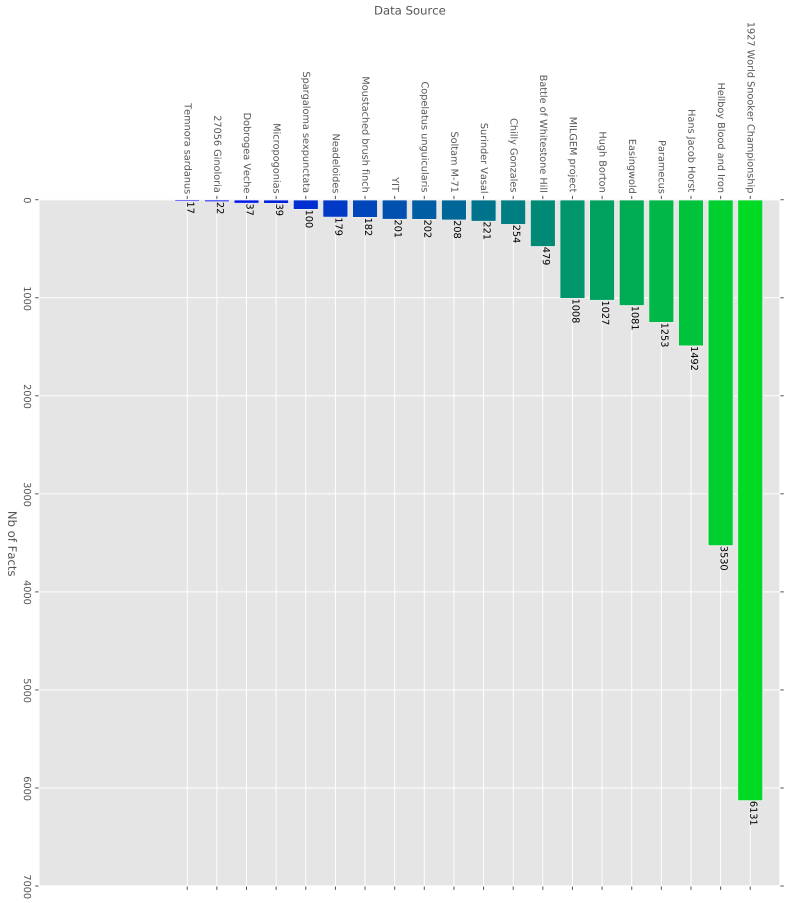


Figure XII.1 : Nombre de triplets extraits du contenu des 10 premières pages rendues par Google Search. Les triplets sont segmentés par entités, tous OIE confondus.

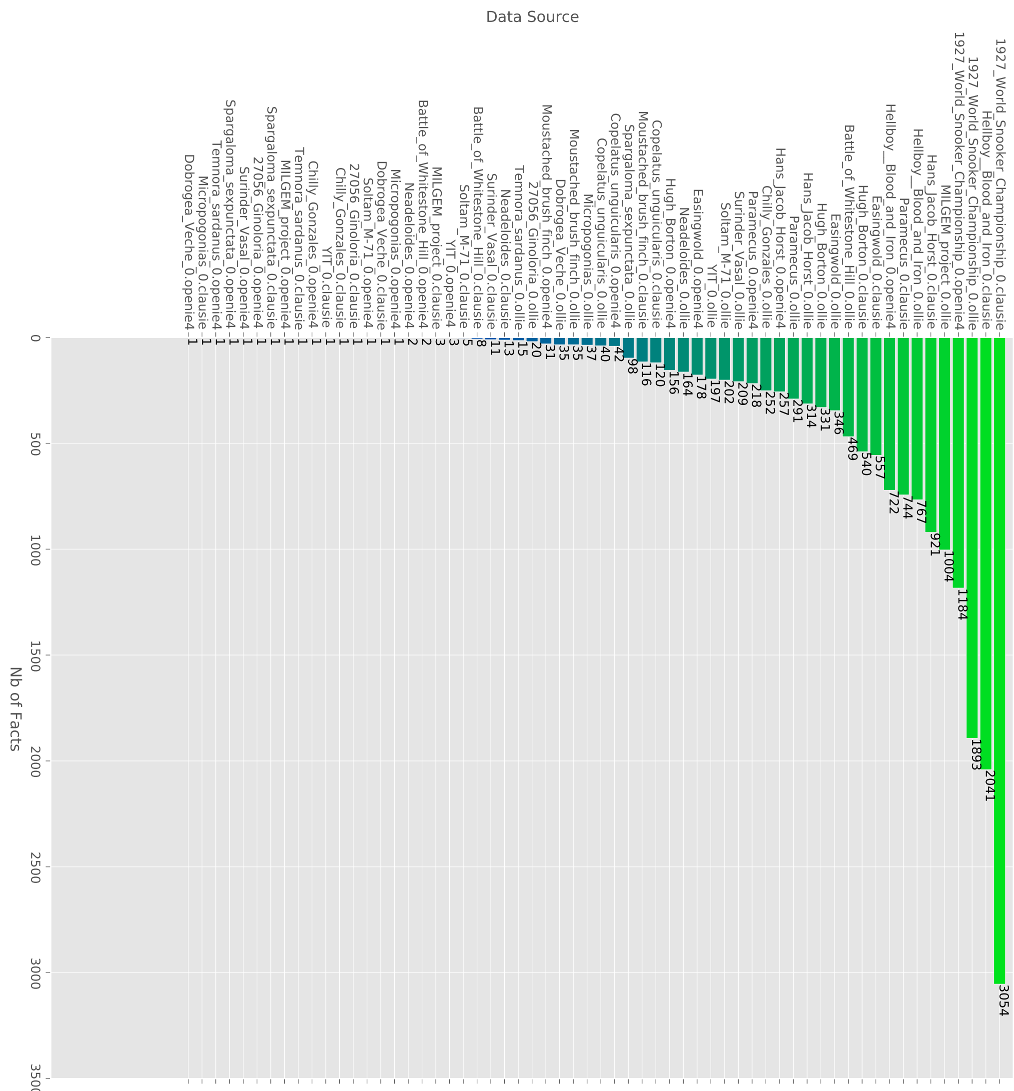


Figure XIII.1 : Nombre de triplets extraits du contenu des 10 premières pages rendues par Google Search. Les triplets sont segmentés par OIE et par entité.