

EXTRACTION AUTOMATIQUE DES QUANTITÉS DE STUPÉFIANTS, DANS LES RÉSUMÉS DE PROCÉDURE ENREGISTRÉS PAR LES FORCES DE SÉCURITÉ INTÉRIEURE, À L'AIDE DE MODÈLES LÉGERS

Mathias ROBERT (*)

(*) SSMSI, Bureau des études et statistiques sur la criminalité organisée

mathias.robert@interieur.gouv.fr

Mots-clés : Analyse textuelle, machine learning, extraction, modèles de langage, entités nommées, stupéfiants.

Domaines : Science des Données – Analyse du Langage Naturel.

Résumé

Le SSMSI (Service statistique ministériel de la sécurité intérieure) conduit des travaux pour étudier l'ensemble des champs infractionnels, notamment ceux de la criminalité organisée, dont les infractions à la législation sur les stupéfiants (ILS), principalement le trafic et l'usage de stupéfiants. Le SSMSI dispose de données administratives issues des logiciels de rédaction des procédures (LRP) de la police et de la gendarmerie nationales (LRPPN, LRPGN). Ces logiciels alimentent des puits de données qui fournissent diverses informations générales sur les mis en cause, les victimes et les circonstances des faits commis. À partir des données brutes issues des puits sont constituées des bases statistiques, facilement exploitables par les chargés d'études. Le présent travail a vocation à enrichir les données sur les ILS disponibles dans ces bases, notamment par l'ajout de variables indiquant le ou les types de stupéfiants en cause, ainsi que leur quantité associée, généralement indiquée sous la forme d'un poids en grammes.

Au sein des puits de données, plusieurs sources permettent d'identifier les stupéfiants : la base des procès-verbaux électroniques, la base OSIRIS gérée par l'Office anti-stupéfiants (OFAST) et la base des « objets » issue de LRPPN. Dans ces bases, quand elle est présente, l'information sur le type et la quantité est inscrite dans des champs spécifiques et est donc facile à extraire. Cependant, ces informations ne sont pas renseignées pour tous les mis en cause pour ILS ; quand le type de stupéfiant et la quantité associée sont inconnus, l'étude des résumés de procédure, petits textes rédigés par les policiers et gendarmes décrivant les manières d'opérer (« *manop* »), devient indispensable. Ces résumés de procédure sont de qualité inégale et ne sont pas toujours remplis.

Les *manop* issues de LRP concernant les mis en cause pour trafic ou usage de stupéfiants mentionnent dans 80 % des cas le(s) type(s) de stupéfiants concerné(s) ; moins de 5 % des *manop* sont vides. Il s'agit le plus souvent de résine de cannabis, d'herbe de cannabis, de cocaïne, d'héroïne et d'ecstasy. À l'aide d'expressions régulières et de listes de mots-clés, il est aisé de détecter quels types de stupéfiants sont impliqués. En revanche, il est plus difficile de repérer la quantité associée. Il s'agit en effet d'identifier un nombre, une unité de mesure et le stupéfiant associé.

Une approche naïve peut consister en l'identification de *patterns* décrivant des poids (« 30 g », « 1,2 kilo », « 250grm », etc.) à l'aide d'expressions régulières. La proximité des différents poids avec les types de stupéfiants détectés dans la manière d'opérer permet de déterminer le stupéfiant le plus probable auquel ce poids appartient. Le cas où autant de stupéfiants que de poids sont détectés est le plus simple. En revanche, pour M types de stupéfiants et N poids, l'affectation est plus complexe à résoudre et s'apparente à un problème d'affectation linéaire, où le coût d'une affectation correspond à la distance dans le texte entre un stupéfiant et sa quantité associée supposée. Pour 10 stupéfiants et 9 quantités, il y a ainsi $\binom{10}{9} \times 9!$ (soit plus de 3 millions) affectations différentes possibles. L'algorithme hongrois permet notamment de résoudre ce problème en temps polynomial mais de nombreuses autres méthodes existent, dont celle de R. Jonker et T. Volgenant.

Puisqu'elle est basée sur des expressions régulières, cette méthode est peu flexible. Il y a de nombreux cas où la quantité de stupéfiant ne saurait être détectée par des expressions régulières (« trente grammes », « 120 de cannabis », « cocaïne : 3+5g », etc.), ou au contraire identifiée fa-lacieusement (« 30kg de légumes », « véhicule immatriculé ab 230 gr », « poids total de 850 g », etc.). Le présent travail présente ainsi différentes méthodes de traitement automatique du langage naturel, permettant l'extraction de quantités dans les textes et des substances auxquelles elles sont liées. Les méthodes présentées ici se veulent légères et adaptées à des capacités computationnelles limitées qui sont celles mises à disposition dans l'environnement professionnel du SSMSI.

Ce problème s'apparente notamment à une situation de reconnaissance d'entités nommées (*Named-Entity Recognition*, NER). Les modèles de NER permettent de reconnaître certaines entités prédéfinies comme des prénoms ou des lieux par exemple. Il est possible d'entraîner un modèle NER sur la détection de l'entité « quantité de résine de cannabis » ou d'une autre substance que l'on sait présente dans une *manop*. L'entraînement du modèle, nécessitant par ailleurs une étape de labellisation manuelle, est une tâche relativement lourde en ressources, c'est pourquoi il est possible de se tourner vers un modèle tel que GLiNER (*Generalist and Lightweight model for Named-Entity Recognition*). Ce dernier permet en « *zero-shot* », soit sans entraînement, de détecter directement l'entité choisie – néanmoins, entraîner un tel modèle reste avantageux pour ses performances.

Ce problème peut également être résolu en mettant à profit les performances des modèles de langage. Si les grands modèles de langage (LLM) sont davantage performants, l'utilisation de plus petits modèles de langage (SLM) de 100 millions à 1 milliard de paramètres est souvent suffisante pour des tâches simples comme l'extraction de quantités, que ce soit en *zero-shot* ou bien après un *fine-tuning*, plus coûteux mais qui donne généralement de meilleurs résultats.

Les analyses montrent enfin que l'utilisation de modèles hybrides, utilisant comme base l'approche naïve mais corrigeant ses défauts à l'aide de SLM, permettent d'obtenir une efficacité supérieure à l'usage brut de modèles tels que les NER et les SLM seuls qui montrent des résultats limités en l'absence de capacités computationnelles importantes.

Abstract

SSMSI, Ministerial Statistical Service for Internal Security (*Service Statistique Ministériel de la Sécurité Intérieure*, SSMSI), conducts works to study a significant number of infraction fields, among which those related to organised crime, including drug-related offenses, mainly drug trafficking and use. The SSMSI has access to administrative data from the procedure drafting software, called *LRP*, used by the national police (LRPPN) and gendarmerie (LRPGN). This software feeds into data pools that provide various general information on suspects, victims, and the circumstances of the offenses committed. The raw data from the data pools is used to create statistical databases that are easily accessible to researchers. The purpose of this work is to enrich the data on drug offenses available in these databases, in particular by adding variables indicating the type(s) of drugs involved and

their associated quantity, generally indicated in grams.

Within the data wells, several sources can be used to identify narcotics: the issued electronic tickets database, the “Osiris” database managed by the Anti-Narcotics Office (*OFAST*), and the “Objects” database from LRPPN. In these databases, when available, information on the type and quantity is entered in specific fields and is therefore easy to extract. However, this information is not provided for all defendants charged with drug offenses; when the type of narcotic and the associated quantity are unknown, it becomes essential to study the procedural summaries, short texts written by police officers and gendarmes describing the modus operandi *manières d’opérer*, “*manop*”. These procedural summaries are of varying quality and are not always completed.

The modus operandi derived from LRP concerning defendants charged with drug trafficking or use mentions the type(s) of drugs involved in 80 % of cases; less than 5 % of modus operandi are blank. The most common drugs are cannabis resin, cannabis herb, cocaine, heroin, and ecstasy. Using regular expressions and keyword lists, it is easy to detect which types of narcotics are involved. However, it is more difficult to identify the quantity involved. This involves identifying a number, a unit of measurement, and the associated narcotic.

A naive approach may consist of identifying patterns describing weights (“30 g,” “1.2 kg,” “250 g,” etc.) using regular expressions. The proximity of the different weights to the types of narcotics detected in the modus operandi makes it possible to determine the most likely narcotic to which this weight belongs. The simplest case is when as many narcotics as weights are detected. However, for M types of narcotics and N weights, the assignment is more complex to solve and resembles a linear assignment problem, where the cost of an assignment corresponds to the distance in the text between a narcotic and its associated supposed quantity. For 10 narcotics and 9 quantities, there are thus $\binom{10}{9} \times 9!$ (i.e., more than 3 million) different possible assignments. The Hungarian algorithm can solve this problem in polynomial time, but many other methods exist, including that of R. Jonker and T. Volgenant.

Since it is based on regular expressions, this method is not very flexible. There are many cases where the quantity of narcotics cannot be detected by regular expressions (“thirty grams,” “120 grams of cannabis,” “cocaine: 3+5g,” etc.), or, conversely, can be identified incorrectly (“30kg of vegetables,” “vehicle registered ab 230 gr,” “total weight of 850g,” etc.). This paper presents various methods of automatic natural language processing that enable the extraction of quantities from texts and the substances to which they relate. The methods presented here are designed to be lightweight and adapted to the limited computational capabilities available in the bureau’s professional environment.

This problem is similar to named entity recognition (NER). NER models can recognize certain predefined entities such as first names or places, for example. It is possible to train a NER model to detect the entity “amount of cannabis resin” or another substance known to be present in a modus operandi. Training the model, which also requires a manual labeling step, is a relatively resource-intensive task, which is why it is possible to choose instead a model such as GLiNER (Generalist and Lightweight model for Named-Entity Recognition). The latter allows “zero-shot” detection, i.e., without training, of the chosen entity—however, training such a model remains advantageous for its performance.

This problem can also be solved by leveraging the performance of language models. While large language models (LLMs) are more powerful, smaller language models (SLMs) with 100 million to 1 billion parameters are often sufficient for simple tasks such as quantity extraction, whether in zero-shot or after fine-tuning, which is more costly but generally yields better results.

The analysis finally shows that the use of hybrid models, on the basis of a naive approach but fixing its drawbacks with the use of SLMs, yields a higher efficiency than the raw use of models such as NER and SLMs alone that show limited results in the absence of high computational capacities.

1 Contextualisation

1.1 Un objectif : la création d'une base dédiée aux stupéfiants en cause

Le SSMSI diffuse des informations relatives aux mis en cause pour usage et trafic de stupéfiants par la police et la gendarmerie nationales, en termes de volume, de caractéristiques sociodémographiques des mis en cause (âge, sexe et nationalité, principalement) et de répartition géographique (jusqu'au niveau communal), depuis 2016. En revanche, à ce jour il n'existait aucune information publiée sur les types de stupéfiants en cause ni leur quantité.

Cette distinction est pourtant primordiale dans l'analyse de ce champ délictuel. Dans une publication à venir, les résultats des analyses montrent que l'âge, la part de femmes et la part d'étrangers des mis en cause sont très différents en fonction du stupéfiant en question ; il en est de même pour leur répartition géographique (Robert, à paraître). Selon le type de stupéfiant concerné, le nombre de mis en cause passe de quelques unités à plusieurs dizaines de milliers (voire centaines de milliers, pour l'usage).

Il était donc opportun pour le SSMSI de lancer des travaux à ce sujet, afin de décrire les dynamiques différentes qui opèrent en fonction du stupéfiant concerné. Pour ce faire, une base permettant d'accéder rapidement aux stupéfiants associés à chacune des procédures d'ILS est nécessaire.

Les informations prioritaires souhaitées sont le type et la quantité de stupéfiant. Par souci de comparabilité, la quantité attendue est toujours un poids en grammes. Ainsi, la base souhaitée est construite selon le schéma suivant :

Numéro de procédure	Type de stupéfiant	Quantité de stupéfiant
X	Cannabis	45 g
Y	Héroïne	7.5 g
Y	Cocaïne	3 g
Z	Cannabis	1 500 g

TAB. 1 – Aspect de la base attendue

Il est possible pour une même procédure d'impliquer un ou plusieurs mis en cause pour l'usage ou le trafic de plusieurs stupéfiants à la fois. Idéalement, il faudrait constituer une base associant, à chaque procédure et chaque mis en cause, les stupéfiants et les quantités de stupéfiants qui le concernent. Comme nous le verrons en section 1.3, ce n'est cependant pas possible, nous ne pouvons disposer, dans chaque procédure, que de la liste des personnes mises en cause dans cette procédure d'une part, et des stupéfiants impliquées dans la procédure d'autre part, sans pouvoir établir de liens entre eux.

1.2 Les sources de données

L'information sur le type et la quantité de produit stupéfiant en cause est renseignée dans plusieurs bases différentes. Afin d'être en mesure de détecter le type et la quantité de stupéfiant pour un maximum de procédures, un ensemble de quatre sources est utilisé.

1.2.1 Les sources primaires : procès-verbaux électroniques, Osiris, Objet

Les trois premières sources sont des bases dans lesquelles l'information sur le type et la quantité de stupéfiant sont écrites en clair, dans des variables dédiées. L'extraction de l'information se

fait donc de manière relativement directe.

La base des procès-verbaux électroniques (PVE), fournie par l'Agence Nationale du Traitement Automatisé des Infractions (ANTAI), contient des informations relatives aux amendes forfaitaires délictuelles (AFD). Les AFD sont un dispositif simplifiant les procédures liées à certaines infractions simples et caractérisées. Depuis 2020, il est possible pour les services de sécurité intérieure de délivrer une AFD pour l'usage de stupéfiants, aux conditions suivantes : le mis en cause doit être majeur, sans spécification de nationalité mais parlant et comprenant le français sans difficulté, possiblement récidiviste mais pas « notoirement connu » pour des infractions à la législation sur les stupéfiants, disposer d'une adresse, ne pas être sous tutelle, curatelle, ou dépositaire de l'autorité publique, être capable de justifier son identité, physiquement présent – c'est-à-dire, pas en fuite –, en présence du produit stupéfiant, ne pas être au volant d'un véhicule à moteur, ne pas faire l'objet de circonstances aggravantes, reconnaître l'infraction, accepter de signer le PVE et de voir ses biens en rapport avec l'infraction confisqués.

En sus, le type de produit stupéfiant et sa quantité doivent être connus par les services de sécurité intérieure de manière certaine, il ne peut s'agir que de cannabis, cocaïne ou ecstasy-MDMA, dans des quantités respectivement inférieures à 50 grammes, 5 grammes, et 5 grammes ou 5 cachets, en l'absence de produits stupéfiants multiples mais possiblement avec plusieurs formes d'un même produit (herbe et résine de cannabis, par exemple).

Si toutes ces conditions sont réunies, il est possible – mais pas systématique – pour les services de sécurité intérieure de délivrer une simple amende forfaitaire, ce qui simplifie les procédures et contribue à la désaturation des tribunaux. L'amende nominale s'élève à 200 euros.

En 2024, deux tiers des mis en cause pour usage de stupéfiants ont reçu une amende forfaitaire délictuelle [1]. Le champ d'étude, regroupant l'ensemble des mis en cause pour usage et pour trafic de stupéfiant (en 2024, respectivement de l'ordre de 300 000 et 50 000 mis en cause), est donc couvert en majorité par des mis en cause par AFD en 2024. Les informations contenues dans cette base sont donc très utiles puisqu'elles fournissent une information pour plus de la moitié des mis en cause en 2024 et une proportion importante sur la période 2016-2023. En outre, ces informations peuvent être qualifiées de fiables, puisque la condition d'absence de doute sur le produit et la quantité en cause conduit à un remplissage exhaustif de ces deux variables dans la base des PVE.

Quand un individu est mis en cause *via* une AFD, la base PVE est ainsi la seule source nécessaire. Cependant, cette source est totalement inutile pour les mis en cause pour usage de stupéfiants hors AFD, ou pour les mis en cause pour trafic de stupéfiants. D'autres bases sont alors mises à profit.

La base Osiris (**O**util et **S**ystème d'**I**nformations **R**elatives aux **I**nfractions sur les **S**tupéfiants), gérée par l'Office Anti-Stupéfiants (OFAST), anciennement OCRTIS, existe depuis 2006 et recense de nombreuses informations sur les auteurs d'infraction à la législation sur les stupéfiants. Elle est notamment alimentée par les données issues du système d'information du Traitement des Antécédents Judiciaires (TAJ), des douanes et des traitements spécifiques de la préfecture de police.

Dans ces bases, une variable indiquant le type de produit comprend plus d'un millier de modalités différentes, qui ne sont pas toutes classées comme produits stupéfiants au sens de l'arrêté du 22 février 1990 fixant la liste des substances classées comme stupéfiants ; il peut également s'agir d'autres substances contrôlées comme les stéroïdes anabolisants, les médicaments sur listes I et II, le protoxyde d'azote...

Les quantités associées sont indiquées dans diverses unités de mesure, qui peuvent être des grammes, milligrammes, kilogrammes ou tonnes, ce qui est facile à convertir en grammes, ou parfois dans des unités de mesure difficilement comparables à un gramme : centilitres (pour une forme liquide), mètres (pour la taille d'un pied de cannabis). Enfin, certaines « unités de mesure » correspondent à des récipients ou des objets contenant la substance stupéfiante, et dans ce cas aucun

poids n'est fourni. La quantité du produit stupéfiant peut ainsi être fournie en nombre de sachets, flacons, gélules, bonbonnes, boulettes, joints... En l'absence de table d'équivalence fournissant le poids des stupéfiants contenus dans un sachet ou un flacon, ces unités de mesure sont ignorées et seules celles qui sont facilement convertibles en grammes sont considérées. Cela a pour effet de baisser le taux de détection réel du poids de stupéfiants, puisque ces quantités sont retirées.

La police nationale est à l'origine d'une autre base, la base des « Objets » directement intégrée dans le logiciel de rédaction des procédures de la police nationale (LRPPN). Elle n'existe pas dans l'équivalent pour la gendarmerie nationale (LRPGN), ainsi les données obtenues ne concernent que les mis en cause pour usage ou trafic de stupéfiants par la police nationale uniquement. Les informations sur la quantité de stupéfiants sont inscrites de la même manière que pour la base Osiris, en revanche pour le type de stupéfiant, si les modalités du type de produit sont les mêmes que pour la base Osiris, il existe une variable supplémentaire indiquant la « marque » de l'objet, qui est un champ libre dans lequel il est possible de renseigner n'importe quelle précision sur le type de stupéfiant. Les marques qui apparaissent le plus souvent, qui sont en réalité simplement des précisions sur le type de stupéfiant en cause dans la plupart des cas, sont exploitées afin d'affiner le type de stupéfiant en cause.

Ces trois bases permettent la détection d'un type de stupéfiant pour 85 % des mis en cause sur la période 2016-2024 (91 % en 2024). Il peut arriver que les sources se contredisent légèrement sur le type de stupéfiant en cause, ou sa quantité. Dans ce cas, la base Osiris est considérée comme une base prioritaire vis-à-vis de la base des Objets : en cas de conflit sur le stupéfiant en cause, la base Osiris sert de référence. Quand ni la base des PVE, ni la base Osiris, ni la base des Objets fournit l'information sur le type et la quantité de stupéfiants, une dernière base peut être mise à profit : la base des résumés de procédure.

1.2.2 Les résumés de procédure, source secondaire indispensable mais inégale

Issus du logiciel de rédaction des procédures de la police et de la gendarmerie nationales (LRPPN - LRPGN), les résumés de procédure sont de petits textes rédigés par les services de sécurité intérieure et qui décrivent les manières d'opérer – souvent abrégées **manop** –, c'est-à-dire, les circonstances des faits commis.

Par conséquent, l'extraction du type de stupéfiant et de sa quantité associée s'y fait par analyse textuelle uniquement. L'extraction du type de stupéfiant ne peut pas toujours être réalisée : parfois, la manière d'opérer n'est pas remplie, parfois, elle ne contient aucune information qui nous soit utile. Néanmoins, l'utilisation des résumés de procédure permettent de passer d'un taux de détection du type de stupéfiant de 85 % sur la période 2016-2024 à 96 %. En 2024 uniquement, il permet de passer d'un taux de détection de 91 % à 98 %.

L'augmentation du taux de détection en 2024 peut notamment s'expliquer par une forte augmentation des AFD délivrées continue depuis leur mise en place en 2020, qui sont renseignées de manière exhaustive dans la base des PVE en ce qui concerne le type et la quantité de produit stupéfiant.

Le traitement de la base des résumés de procédure est inutile dans le cadre des AFD, puisque l'information est toujours fournie dans la base PVE. Dans les autres cas (c'est-à-dire, dans les cas d'usage de stupéfiants sans AFD, ou les cas de trafic de stupéfiants), le type de stupéfiant est non détecté à des niveaux stables (6 % des mis en cause aussi bien sur la période 2016-2024 qu'en 2024 uniquement).

La répartition globale de la détection par source sur la période 2016-2024 (FIG. 1) montre que la base des Objets et la base Osiris seules ne contribuent respectivement qu'à 1 % et 2 % de la

détection, et qu'en réalité presque la moitié des mis en cause ont au moins un stupéfiant détectable dans chacune des trois sources. En revanche, 16 % des mis en cause dont le type de stupéfiant a pu être extrait n'ont pu l'être seulement que par la base des résumés de procédure.

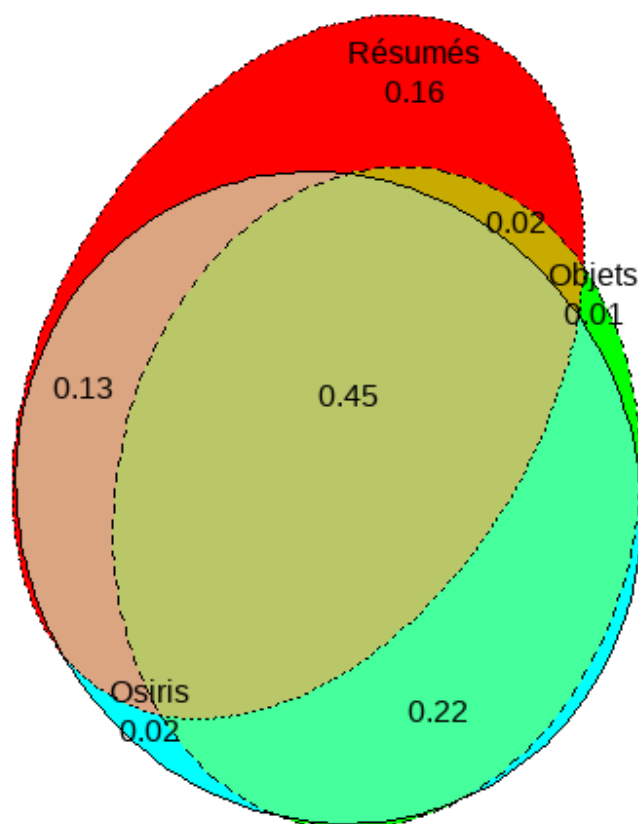


FIG. 1 – *Détection des stupéfiants par source sur la période 2016-2024, hors AFD*
Lecture : 45 % des mis en cause hors AFD ont un stupéfiant détecté à la fois dans les bases Osiris, Objets et les résumés de procédure sur la période 2016-2024

1.3 La construction de la base

L'objectif est de produire une base recensant, pour chaque procédure d'ILS et idéalement chaque mis en cause de ces procédures, les stupéfiants impliqués dans la procédure et les quantités associées.

Pour construire la base, il faut donc pour chaque ligne pouvoir déterminer à la fois le type de stupéfiant, et sa quantité associée.

1.3.1 L'utilisation de la procédure à défaut du mis en cause pour l'indexation

L'idée initiale de la base est d'obtenir les différents types de stupéfiants associés à chaque mis en cause. Cependant, les sources qui sont utilisées n'utilisent pas le mis en cause comme unité de compte ; en d'autres termes, les stupéfiants n'y sont pas enregistrés vis-à-vis du mis en cause, mais vis-à-vis d'une unité de compte plus large : ou bien de la procédure tout entière, ou bien du fait. Seule la base des procès-verbaux électroniques ne pose pas de problème, puisqu'en raison des conditions de délivrance d'une AFD, une procédure est associée à un seul mis en cause et à un seul fait – ainsi joindre la base des PVE à la base des mis en cause sur les numéros de procédure est équivalent à la joindre sur les numéros d'identifiant de mis en cause.

Pour les résumés de procédure et Osiris, les stupéfiants sont reliés à une procédure. Pour la police et la gendarmerie, une procédure correspond à l'ensemble des procès-verbaux établis par les services de sécurité intérieure à la suite de la constatation d'une infraction. Une procédure peut contenir une ou plusieurs infractions. Cette notion se rapproche de celle de l'affaire utilisée par les parquets¹. Ainsi, une même procédure peut être reliée à plusieurs mis en cause.

De même, pour la base des Objets, les stupéfiants étant reliés à un fait, donc à une infraction ou un ensemble d'infractions au sein d'une procédure, il peut aussi être associé à plusieurs mis en cause.

Par conséquent, puisqu'il n'est pas possible d'obtenir l'information dans les sources sur les stupéfiants en cause individu par individu, cela ne sera pas non plus possible dans la base finale ; l'information est alors agrégée au niveau de la procédure.

1.3.2 Les cas simples : la base des PVE, Osiris et des Objets

Dans la plupart des cas, la source principale est la base des PVE, la base Osiris ou des Objets. Dans l'intégralité des cas d'usage de stupéfiants avec AFD délivrée, la base des PVE fournit, par des variables simples à exploiter, le type de stupéfiant et la quantité associée.

Dans le cas contraire, quand le type de stupéfiant peut être extrait, c'est dans 84 % des cas grâce à la base Osiris ou la base des Objets (cf. figure 1) – et très souvent, ce sont en réalité les deux bases qui sont capables de fournir l'information. Dans une majorité des cas donc, l'extraction du type de stupéfiant et de leur quantité associée se fait assez facilement grâce à des variables distinctes donnant le type et la quantité de stupéfiants, quitte à regrouper les types de stupéfiants selon les agrégats qui conviennent à la classification réalisée par le SSMSI.

En revanche, dans les 16 % de cas restants, ce sont les résumés de procédure qui fournissent l'information et cela requiert l'extraction du type et de la quantité de stupéfiant par des méthodes plus complexes d'analyse textuelle.

1.3.3 Un aperçu des résumés de procédure

Si un résumé de procédure sert à expliciter les circonstances des faits commis, il n'est pas toujours rédigé de manière détaillée.

Pour les cas d'usage ou de trafic de stupéfiants, il comprend, dans plus de 90 % des cas, entre 100 et 1 000 caractères, et entre 20 et 200 mots.

En voici quelques exemples relativement courts (l'ensemble des majuscules et des accents ont été supprimés, certaines informations sensibles sont anonymisées) :

1. au cours d'une perquisition realisee dans le cadre d'une enquete preliminaire, nous decouvrons des produits stupefiants au domicile de l'interesse. il est ainsi decouvert 17,67g de resine de cannabis et 16,53g d'herbe de cannabis.
2. 44,9 g de cannabis etait decouvert dans le vehicule de [INDIVIDU]. il etait place en gav. entendu, il declarait que les stupefiants ne lui appartenaient pas.

1. Voir le glossaire disponible sur le site Interstats.

3. un trafic de stupefiants est demantele sur le secteur de [COMMUNE], portant sur du cannabis et sur de la cocaine. tete de reseau et consommateurs sont entendus.

Dans les deux premiers cas, les types de stupéfiants et les quantités associées sont mentionnés.

Dans le dernier cas, les quantités en cause ne semblent pas être connues et ne sont donc pas mentionnées.

1.3.4 L'extraction du type de stupéfiant dans les résumés de procédure

Dans les résumés de procédure, les styles d'écriture sont relativement diversifiés et ainsi, un même stupéfiant peut avoir plusieurs dénominations différentes :

4. le mis en cause est trouve porteur d'un couteau, une balance de precision et d'un pochon contenant 0.6 gramme de resine de cannabis.
5. lors d'une perquisition dans le cadre d'une enquete de flagrance concernant un vol avec violence, decouvrons au domicile de l'interesse, 0.7 grammes de haschich et des objets lies a la consommation de stupefiants.
6. le nomme [INDIVIDU] etait interpelle avec 5.2 gr de rc et un couteau.audtionne, il reconnaissait les faits.sur les instructions du parquet, mesure de protection judiciaire de la jeunesse demandee .
7. le [DATE] a [HEURE], l'auteur est controle a la piscine municipale de [COMMUNE] par la police municipale. il est porteur de 2 barrettes de shit.
8. [INDIVIDU] etait interpelle avec 02 grammes de rsine de cannabis. procedure simplifiee. destruction du produit ulterireurement par moyen adapte.
9. le conducteur est controle en barriere de peage [NOM DE L'AUTOROUTE] sur la commune de [COMMUNE]. sur requisition pr [COMMUNE], nous procedons a la fouille du vehicule. le chien decouvre une boulette de hachiche de 0.6 grammes. le conducteur nous declare spontanement etre consommateur occasionnel et avoir achete cette boulette pour le week end.

Dans tous les cas présentés ici, le stupéfiant en cause est la résine de cannabis. Cependant, on la trouve dans le résumé de procédure *via* les termes « résine de cannabis », « haschich », « shit », « rc » ou encore des versions de ces termes contenant des fautes de frappe comme « rsine de cannabis » ou « hachiche », qui pourraient entraver la détection : un stupéfiant peut être qualifié par plusieurs mots différents et il peut y avoir des fautes de frappe. La flexibilité est donc importante dans l'optique de détecter un maximum de variantes différentes.

Une liste de mots-clés couvrant la plupart des dénominations des stupéfiants et des fautes de frappe courantes permet d'obtenir un taux d'extraction proche de l'optimum avec une utilisation simple d'expressions régulières.

En revanche, pour l'exploitation ultérieure de la base, le codage des types de stupéfiants revêt une importance ; ainsi on veut notamment regrouper ensemble les termes « résine de cannabis » et « haschich » sous une même dénomination de résine de cannabis, mais on veut aussi dans certains cas pouvoir les agréger avec les autres formes de cannabis tels que l'herbe ou les pieds de cannabis, tout en permettant de pouvoir traiter les différentes formes de cannabis séparément. Il en est de même pour tous les types de stupéfiants.

Cela requiert d'utiliser une certaine hiérarchie à deux niveaux, au sein de laquelle la résine et l'herbe de cannabis seraient séparées au deuxième niveau mais regroupées au sein du premier niveau. L'Office des Nations unies contre la drogue et le crime (ONUDC) fournit une nomenclature, terminologie des drogues [2] ainsi formée. Elle sert de base à la nomenclature utilisée par le SSMSI. Quelques ajouts ont été effectués afin d'inclure des stupéfiants qui ne l'étaient pas, comme la kétamine.

Ainsi, avec cette terminologie, chaque stupéfiant est codé avec une combinaison d'un chiffre et d'une lettre, le premier classant le stupéfiant dans des catégories assez larges (1 : cannabis, 4 : opioïdes, 8 : hallucinogènes...) et la seconde précisant le sous-type de stupéfiant (1A : résine de cannabis, 4B : méthadone, 8A : LSD...). L'ensemble de la classification est disponible en annexe.

1.3.5 Le cas difficile de l'extraction de la quantité associée dans les résumés

Dans les résumés de procédure, il n'est pas aussi simple de détecter la quantité que le type de stupéfiant. En effet, l'extraction de la quantité est soumise à la détection de trois éléments :

- une valeur numérique ;
- une unité de mesure ;
- le stupéfiant auquel elle est associée.

Si à première vue les quantités semblent, contrairement aux types de stupéfiants, être construites de manière relativement formalisée, sous la forme « XXX [unité de mesure, généralement grammes] de [type de stupéfiant] » – comme par exemple : « 15 grammes d'herbe de cannabis » –, il n'en est rien, et en réalité de nombreux cas symptomatiques d'extraction difficile des quantités apparaissent :

10. [INDIVIDU] a pris la fuite a la vue des policiers, lors du controle il s'est debarasse de deux sachets contenant de la resine de cannabis (5 grammes) [INDIVIDU] entendu reconnait l'usage de produits stupefiants. copj mineur
11. la mec² est controlee en possession de 0.2 de ketamine en poudre, de 1g de cannabis et de 3 couteaux suisses en train de fumer sur la parking de la gare.
12. le pere du suspect decouvre dans sa chambre des produits stupefiants. la perquisition menee permet la saisie de quatorze virgule soixante-quatorze grammes de resine de haschich.

Ainsi il est tout à fait possible de rencontrer des quantités mentionnées après le stupéfiant, par exemple entre parenthèses (cas 10.) ; l'absence de formalisation de la quantité dans les résumés de procédure induit que ces trois informations ne se suivent pas forcément dans cet ordre. Si la valeur numérique précède quasiment toujours l'unité de mesure, le stupéfiant associé peut être placé devant ou derrière la quantité. Dans le cas où le résumé de procédure comporte plusieurs stupéfiants différents, cela peut poser des problèmes dans la tâche de détection.

L'unité de mesure peut aussi être une source de confusion. Il est ainsi possible de rencontrer des quantités avec omission de l'unité de mesure (cas 11.), ou encore des quantités écrites en lettres (cas 12.). Ce genre de cas peut compliquer l'extraction pour certains algorithmes simples.

Par ailleurs, il est courant dans les résumés de procédure de nommer un individu, qu'il soit mis en cause, victime ou témoin, par ses initiales. Ainsi, les motifs « gr » ou « kg » par exemple peuvent

²MEC : mis(e) en cause

être à la fois des unités de mesure ou bien des initiales. Ainsi le motif « A 15 h 12 kg est interpellé » peut facilement être interprété à tort comme un poids de stupéfiants alors qu'il désigne l'individu nommé K. G.

Quand bien même une valeur numérique et une unité de mesure sont correctement associées, rien n'indique qu'il s'agisse du poids d'un type de stupéfiants ; il peut s'agir d'un poids de légumes dans un camion qui transporte également des stupéfiants, du poids du total de stupéfiants saisis...

Enfin, même si un poids est extrait correctement, le stupéfiant associé n'est pas toujours simple à identifier notamment dans le cas où il y a plusieurs stupéfiants, et en particulier dans le cas où, pour une raison ou une autre, il n'y a pas le même nombre de quantités que de stupéfiants associées dans le résumé de procédure.

À cause de l'ensemble de ces difficultés, l'extraction de la quantité de chaque stupéfiant n'est absolument pas triviale. Ainsi, plusieurs méthodes sont explorées puis comparées afin de pouvoir extraire cette quantité.

2 L'extraction de la quantité par une méthode simple

L'objectif désormais est donc, en partant d'une liste de stupéfiants détectés dans un résumé de procédure, d'extraire l'ensemble des quantités – en particulier, des **poids** – associées à chacun de ces stupéfiants.

2.1 La construction d'une expression régulière flexible

La première méthode est une méthode que l'on peut qualifier de « naïve », puisqu'il s'agit d'une extraction qui se base uniquement sur des motifs à détecter dans le texte, sans aucune flexibilité sur des cas non prévus de formulations inhabituelles. Elle est déterministe : contrairement à des méthodes fondées sur des modèles de langage, une extraction de quantités par cet algorithme renverra toujours le même résultat.

Le but de cette méthode est de construire une expression régulière, qui permet de décrire un motif correspondant à des manières usuelles d'écrire une quantité. Ainsi l'expression régulière doit inclure des motifs tels que :

- 3 g
- 3g
- 3,5g
- 03.5 grammes

Par conséquent, elle doit détecter l'ensemble des motifs satisfaisant chacune des conditions suivantes :

- commence par un nombre, de longueur arbitraire
- facultativement, suivi immédiatement par un point ou une virgule, suivi d'un autre nombre
- facultativement, suivi d'une espace
- obligatoirement suivi par une suite d'une ou plusieurs lettres

L'expression régulière associée s'écrit alors :

$$\backslash\mathbf{b}(\backslash\mathbf{d}^+)([.,]\backslash\mathbf{d}^+)?\backslash\mathbf{s}^+([a-zA-Z]^+)\backslash\mathbf{b}$$

Elle retranscrit exactement les conditions précédentes.

Cependant, elle permet d'extraire des nombres suivis d'unités de mesure – voire même, simplement des mots – qui ne sont pas forcément des poids, comme par exemple :

- 14 km
- 14 h
- 14 individus
- « 14.30 nous » (dans une phrase comme « à 14.30 nous procédons au contrôle »)

Pour limiter l'extraction de quantités qui ne sont pas des poids, une liste d'unités de mesure dérivées du gramme (kilogramme, tonne, milligramme...) est dressée. Il est nécessaire qu'elle soit flexible, incluant « g », « gramme », « grammes », « gr », « grs », « grm », « grms », « grs », « gms », ou encore « gm » rien que pour les grammes, car il est possible de trouver toutes ces formes dans les résumés de procédure.

En revanche, cela n'exclut pas quelques cas symptomatiques de détection fallacieuse comme « le véhicule immatriculé ab 210 gr » (interprété comme 210 grammes) ou « à 18.30 kg est interpellé » (interprété comme 18,30 kilogrammes) ; ces cas se révèlent cependant rares.

Cette manière de procéder implique que les quantités non convertibles en grammes (centilitres d'une forme liquide de stupéfiant, mètres pour une taille de pied de cannabis...) et les quantités indiquées par nombre de contenants (flacons, sachets, joints, gélules...) sont supprimées. Il serait pourtant envisageable que les poids de ces contenants soient éventuellement approchés avec des approximations ou des poids moyens issus de travaux externes. Par exemple, l'estimation du poids du cannabis dans un joint a été réalisée dans le cadre de travaux de recherche [3], tout comme le poids moyen des comprimés d'ecstasy [4].

Cependant, d'après les études correspondantes, ces données varient beaucoup selon le contexte d'utilisation, la localisation ou la période – les cachets d'ecstasy étant notamment en moyenne de plus en plus gros au fur et à mesure des années.

Il se montrerait ainsi hasardeux de convertir en grammes ces doses qui pourraient paraître à première vue relativement standard, alors qu'elles sont en réalité soumises à d'importantes variabilités.

2.2 Le problème d'affectation linéaire

Détecter un certain nombre de poids dans un résumé de procédure ne suffit pas : il faut les relier au stupéfiant associé. En résumé, on dispose d'une liste de M types de stupéfiants d'un côté, et de N quantités de l'autre.

Le cas où $M = N$ est le plus simple. En effet, il est assez peu probable que les types de stupéfiants apparaissent dans le résumé de procédure dans un ordre différent que les quantités. Ainsi il fait sens d'associer au premier type de stupéfiant T_1 la première quantité Q_1 , au deuxième type de stupéfiant T_2 la quantité Q_2 , etc.

Le cas où $M \neq N$ est en revanche plus complexe. Il s'agit d'un cas dérivé du *problème d'affectation linéaire*.

Le problème d'affectation linéaire est un problème classique d'optimisation combinatoire. La plupart du temps, il est exprimé comme suit : on dispose de M tâches à réaliser par un ensemble de N employés. Chaque employé ne peut effectuer qu'une tâche. Chaque affectation d'un employé à une tâche représente un coût. Ce coût peut par exemple dépendre des compétences de chaque employé et ainsi, chaque couple d'une tâche et d'un employé a un coût différent. L'objectif est de

trouver quel employé affecter à quelle tâche afin de minimiser le coût total. Dans un cas où il y a plus d'employés que de tâches, certains employés ne se verront pas affecter de tâche. Dans un cas où il y a moins d'employés que de tâches, certaines tâches ne seront pas pourvues.

Ce problème est directement transposable à notre cas : on dispose de M types de stupéfiants que l'on souhaite affecter à N quantités. Le coût d'affecter un type de stupéfiant T_i à une quantité Q_i dans le texte correspond intuitivement à la distance qui sépare le stupéfiant de la quantité : deux termes proches sont plus vraisemblablement liés que deux termes éloignés dans le texte. Pour attribuer à chaque type de stupéfiant une valeur pour coder sa position, le caractère central dans son nom est sélectionné (par exemple, « cocaïne » fait 7 caractères, le caractère central correspond donc au quatrième, « a ») et son index dans le résumé de procédure est retenu. On fait de même pour les quantités. La distance entre les deux termes correspond alors à la différence des deux positions.

Voici un exemple fictif :

20gr de cocaïne sont retrouvés chez le suspect, il nous remet également un pochon de résine de cannabis.

Dans ce cas, les positions sont codées ainsi :

Types de stupéfiants			Quantités		
Type	Terme	Position	Quantité	Terme	Position
T_1	cocaïne	12	Q_1	20gr	2.5
T_2	résine de cannabis	94.5			

Dans ce cas simple il n'y a que deux possibilités d'affectation :

- < Cocaïne ; 20gr >
- < Résine de cannabis ; 20 gr >

Pour cela on calcule les distances avec la différence des positions : dans le premier cas, la distance est de 9.5, dans le deuxième cas elle est de 92. On affecte donc « cocaïne » à la quantité « 20 grammes ». La résine de cannabis n'a pas de quantité associée. C'est effectivement le résultat attendu.

Si le cas présenté ici était simple, il arrive que les cas soient plus complexes, et ainsi par exemple, pour 9 stupéfiants et 10 quantités, il y a $\binom{10}{9} \times 9!$ (soit plus de 3 millions) affectations différentes possibles.

Le problème d'affectation linéaire a fait l'objet de nombreux travaux de recherche et de nombreux algorithmes existent pour sa résolution [5]. La plupart d'entre eux dérivent de l'algorithme de Dijkstra issu du problème du plus court chemin, de manière plus ou moins directe.

L'algorithme classique utilisé pour le problème d'affectation linéaire est l'algorithme hongrois, ou algorithme de Kuhn-Munkres. Il permet de résoudre le problème en temps polynomial – $O(n^4)$. Il a ensuite fait l'objet de nombreuses améliorations.

Dans la librairie Python `scipy`, l'algorithme employé pour résoudre le problème d'affectation linéaire est une version de l'algorithme de R. Jonker et T. Volgenant [6] d'une complexité polynomiale également, mais en $O(n^3)$. C'est celui qui est utilisé pour notre programme d'extraction.

Voici un exemple de résumé de procédure anonymisé.

le [DATE] a [HEURE], agissant sur [CONTEXTE] a [COMMUNE], [ADRESSE], agglomération de [COMMUNE], nous procedons au controle du vehicule de marque [MARQUE], immatricule [IMMATRICULATION]. le vehicule sort de [CONTEXTE]. le passager detient 01 joint, 02 cachets d'ecstasy, 03 pochons de cocaine (0.62 gr), 03 morceaux de resine de cannabis (72.80 grs) et un pochon de ketamine(0.17 gr). il detient sur lui 310 € en numeraire. a son domicile, a [COMMUNE], la perquisition faite par la brigade locale permet la decouverte de 26 cachets d'ecstasy. lors de la fouille judiciaire sur sa personne il est decouvert 13.6 grammes de resine de cannabis, 1.5 gramme de cocaine, 1.5 gramme de mdma et un couteau presentant des traces de resine de cannabis sur la lame.

Ce cas est un exemple d'extraction complexe : on trouve des stupéfiants cités parfois avant les quantités, parfois après, de nombreux stupéfiants ne sont pas associés à des poids utilisables (par exemple : « 1 joint »), et le nombre de stupéfiants cités est important.

Les stupéfiants détectés sont, dans l'ordre : cannabis (« joint »), ecstasy, cocaïne, résine de cannabis, kétamine, ecstasy, résine de cannabis, cocaïne, MDMA, résine de cannabis.

Les poids détectés sont, dans l'ordre : 0.62 gr, 72.80 grs, 0.17 gr, 13.6 grammes, 1.5 gramme, 1.5 gramme.

Il y a donc 10 stupéfiants pour 6 poids, soit plus de 150 000 possibilités d'affectation, dont les positions sont codées ainsi :

Types de stupéfiants			Quantités		
Type	Terme	Position	Quantité	Terme	Position
T_1	joint	291	Q_1	0.62 gr	344
T_2	ecstasy	312	Q_2	72.80 grs	391
T_3	cocaine	335	Q_3	0.17 gr	426
T_4	resine de cannabis	374.5	Q_4	13.6 grammes	664.5
T_5	ketamine	417.5	Q_5	1.5 gramme	698.5
T_6	ecstasy	588	Q_6	1.5 gramme	721.5
T_7	resine de cannabis	683.5			
T_8	cocaine	712			
T_9	mdma	733.5			
T_{10}	resine de cannabis	784.5			

Pour ce résumé de procédure, l'algorithme renvoie que la somme des distances est minimale pour l'affectation suivante : (T_3, Q_1) , (T_4, Q_2) , (T_5, Q_3) , (T_7, Q_4) , (T_8, Q_5) , (T_9, Q_6) . La somme des distances vaut alors 78,5 et se décompose ainsi :

Type	Quantité	Distance
T_3	Q_1	9
T_4	Q_2	16.5
T_5	Q_3	8.5
T_7	Q_4	19
T_8	Q_5	13.5
T_9	Q_6	12
Distance totale		78.5

Cela correspond aux couples types-quantités suivants :

- cannabis : *pas d'affectation*
- ecstasy : *pas d'affectation*
- cocaïne : 0.62 g
- résine de cannabis : 72.8 g
- kétamine : 0.17 g
- ecstasy : *pas d'affectation*
- résine de cannabis : 13.6 g
- cocaïne : 1.5 g
- MDMA : 1.5 g
- résine de cannabis : *pas d'affectation*

Il s'agit exactement du résultat attendu, que l'on aurait pu obtenir par une affectation manuelle coûteuse en temps et en ressources.

Une autre approche qui aurait pu être envisagée, plus simple, serait de raisonner de proche en proche. Dans ce cas, on cherche le couple type-quantité le plus proche (ici, le couple (T_5, Q_3) avec une distance de 8.5); on procède ainsi jusqu'à ce qu'il ne reste plus aucun type de stupéfiant ou quantité non affectée. L'affectation suivant cette méthode donnerait les résultats ci-après :

Type	Quantité	Distance
T_5	Q_3	8.5
T_3	Q_1	9
T_8	Q_6	9.5
T_7	Q_5	15
T_4	Q_2	16.5
T_9	Q_4	69
Distance totale		127.5

Ce raisonnement de proche en proche donne dans cet exemple une distance totale très supérieure à l'allocation précédente, et seule la moitié des affectations sont correctes. Cela s'explique simplement : dans la portion de texte « 13.6 grammes de résine de cannabis, 1.5 gramme de cocaïne, 1.5 gramme de mdma », les types de stupéfiants sont plus proches des quantités qui les suivent que celles qui les précèdent : par exemple, le terme « résine de cannabis » est plus proche de « 1.5 gramme » que de « 13.6 grammes ».

L'utilisation d'un algorithme cherchant à optimiser la distance totale, tel que celui de Jonker-Volgenant, plutôt que de chercher tour à tour le couple type-quantité le plus proche, permet de limiter ces erreurs.

3 Les modèles de *Named Entity Recognition*, NER

Une autre méthode permettant d'extraire les quantités de stupéfiants dans les résumés de procédure repose sur la reconnaissance d'entités nommées (*Named Entity Recognition* – NER).

3.1 Les modèles de reconnaissance d'entités nommées prédéfinies

La reconnaissance d'entités nommées consiste en une tâche d'identification automatique dans un texte segmenté. Chaque segment correspond à une « entité » souhaitée, par exemple, un prénom

ou encore une adresse. Il ne désigne donc pas un modèle en particulier, mais simplement cette tâche d'extraction.

À l'origine, des modèles s'appuyant sur des logiques séquentielles et des règles linguistiques étaient utilisés pour réaliser cette tâche [7].

Depuis plusieurs dizaines d'années, les modèles d'apprentissage profond les ont ensuite peu à peu remplacés, parce qu'ils sont à la fois plus flexibles et plus efficaces. Les premiers modèles de *deep learning* sur le sujet mettent à profit les réseaux de neurones, notamment les architectures à base de *Long Short Term Memory* (LSTM) couplés à des réseaux de neurones convolutifs (CNN) ou bien des champs aléatoires conditionnels (CRF) [8].

L'introduction des *transformers* en 2017 [9], constituent une révolution dans la tâche de NER – et plus globalement dans le milieu du traitement du langage naturel. Contrairement aux architectures séquentielles classiques, les *transformers* s'appuient sur un mécanisme d'attention, qui permet de pondérer l'importance relative de chaque mot par rapport à tous les autres mots d'une phrase. Il permet notamment de gérer la polysémie, c'est-à-dire les différences de sens des mots en fonction du contexte.

Les *transformers* sont composés de couches empilées d'encodeurs et/ou de décodeurs ; cette architecture sert de socle à de multiples modèles modernes comme GPT ou BERT.

Introduit par Google en 2018, BERT (*Bidirectional Encoder Representations from Transformers*) est l'un des premiers modèles à exploiter pleinement l'architecture des *transformers* pour le traitement du langage naturel [10]. En s'appuyant sur un encodeur *transformer* bidirectionnel, BERT modélise simultanément le contexte gauche et droit de chaque mot, améliorant ainsi la compréhension des différentes relations sémantiques dans la phrase.

Il est pré-entraîné sur des tâches de prédiction de mots masqués et de classification de phrases. Pour le NER, une couche de classification spécifique est ajoutée ; elle analyse ces représentations séquence par séquence pour prédire le type d'entité – si la séquence en question est un prénom, une adresse, etc. Cette approche a rapidement supplanté les méthodes classiques notamment grâce à sa capacité à comprendre les ambiguïtés liées au contexte.

La plupart des modèles de NER actuels utilisent l'architecture BERT. S'il est possible de *fine-tuner* le modèle BERT pour qu'il soit en mesure d'extraire les entités prédéfinies de notre choix, cela est coûteux en ressources : en l'absence de capacités computationnelles suffisantes, il est alors possible de se tourner vers des modèles pré-entraînés.

Pour notre cas, les modèles pré-entraînés sur des corpus en français seraient *a priori* plus efficaces. C'est le cas de CamemBERT, un modèle BERT spécifiquement entraîné sur le français pour de nombreuses tâches, notamment du NER. Ses performances sont meilleures sur un corpus en français, et le modèle obtient « de meilleurs F1-scores³ que les architectures basées sur les CRF, aussi bien basés sur des réseaux de neurones ou non, et que des modèles BERT multilingues *fine-tunés* » [11]. Cependant, il est entraîné sur des personnes, des lieux, des objets et des organisations. L'utilisation d'un tel modèle ne permet donc pas de détecter des poids de stupéfiants, en l'état.

Enfin, certains modèles de NER basés sur BERT sont thématiques, et permettent la détection d'entités spécifiques à un champ : on peut notamment citer Legal-BERT pour le droit, capable de détecter dans des corpus juridiques notamment des titres, des lieux ou des lois/juridictions [12] ; on peut également citer les modèles BERT en biologie ou en médecine, comme BioBERT [13] ou pour

3. C'est-à-dire que le modèle de NER parvient à identifier les entités avec à la fois une grande précision (une proportion importante des entités prédites comme positives sont effectivement positives) et un bon rappel/*recall* (une part importante des vrais positifs sont effectivement identifiés par le modèle), traduisant une performance globale élevée.

un modèle en français, CamemBERT-bio [14], capables d'identification d'une large gamme d'entités relatives à la biologie. Ces thématiques sont assez proches de la problématique d'infractions portant sur les stupéfiants pour l'aspect juridique, ou de consommation de stupéfiants pour l'aspect sanitaire et médical. Néanmoins, ils ne sont pas suffisamment spécifiques pour être en mesure d'identifier des quantités de stupéfiants.

3.2 UniversalNER, un modèle de NER généraliste

Cependant, il existe des modèles d'identification d'entités nommées qui soient généralistes, c'est-à-dire pas uniquement capables d'extraire une entité parmi un ensemble d'entités prédéfinies gérées, mais bien toutes sortes d'entités qui peuvent être choisies par l'utilisateur sans limite. UniversalNER en est un exemple [15].

Ce modèle n'est pas basé sur des structures de *transformers* comme BERT, mais sur des grands modèles de langage (LLM). Le modèle utilise une distillation d'un modèle de LLM conversationnel – ChatGPT – afin d'en faire un « modèle étudiant » bien plus léger, adapté à des tâches spécifiques comme celle de l'extraction d'informations ouvertes (c'est-à-dire des entités librement choisies).

Avec une infime fraction de paramètres, UniversalNER acquiert non seulement la capacité de ChatGPT à reconnaître des types d'entités totalement arbitraires, mais il surpasse également sa précision dans la tâche de NER de 7 à 9 points du F1-score en moyenne.

S'il s'agit peut-être du modèle le plus performant dans sa catégorie pour la détection généraliste d'entités nommées, il est basé sur un très grand nombre de paramètres. Il est ainsi très lourd et donc difficilement exploitable par le SSMSI avec des capacités limitées. Il est donc nécessaire pour le service d'exploiter un modèle plus léger.

3.3 GLiNER, un modèle de NER généraliste et léger

Le modèle GLiNER (*Generalist and Lightweight model for Named Entity Recognition*) est un autre exemple de modèle de NER d'identification de toutes sortes d'entités nommées [16].

GLiNER permet de résoudre deux de nos problématiques : la trop grande spécificité des autres modèles de NER sur les types d'entités qui sont identifiables, comme UniversalNER, mais surtout leur coût en ressources (ce modèle étant qualifié de léger, *lightweight*) : les modèles GLiNER sont disponibles avec 50 millions, 90 millions et 300 millions de paramètres, là où UniversalNER fournit des versions à 7 et 13 milliards de paramètres.

Il est par ailleurs entraîné sur des datasets anglophones et multilingues. Le modèle est tout à fait adapté à une utilisation en *zero-shot* (c'est-à-dire, sans *fine-tuning*) aussi bien pour des textes en anglais que dans d'autres langues.

Comme les modèles précédemment présentés – hormis UniversalNER –, GLiNER est un modèle basé sur les *transformers* bidirectionnels dont BERT fait partie. Le modèle prend en entrée un texte ou une phrase, une suite de tokens, etc. (comme peuvent l'être nos résumés de procédure), et une liste d'entités à détecter, qui est entièrement libre (par exemple, il pourrait s'agir de la liste suivante: ['stupéfiant', 'quantité de stupéfiant'] qui correspondrait à nos besoins).

Chaque token est représenté par le *transformer* bidirectionnel ; les représentations des entités sont transmises au réseau, tandis que les représentations des mots en entrée sont transmises à une couche de représentation de *segments* (« *spans* ») composés d'un ou plusieurs mots, afin de calculer les représentations pour chaque segment. Enfin, un score de correspondance est calculé entre les représentations d'entités et les représentations de segments.

Dans l'exemple de la figure 2, tiré de l'article relatif au modèle GLiNER [16], le segment (0, 1), correspondant à « Alain Farley » a un score de correspondance de 0.9 avec l'entité « personne »,

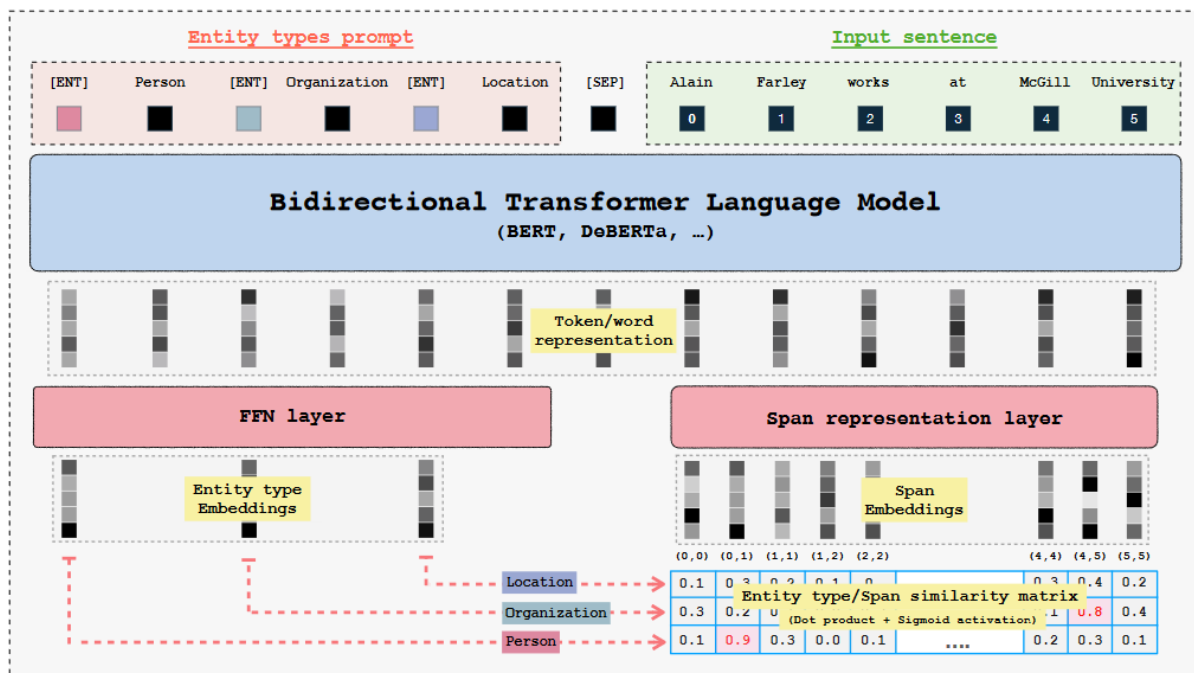


FIG. 2 – Architecture du modèle GLiNER, tiré de l'article présentant le modèle [16]

alors qu'il est seulement de 0.2 avec l'entité « organisation » et 0.3 pour l'entité « localisation ». Il y a donc de grandes chances qu'Alain Farley désigne le nom d'une personne, plus qu'une organisation ou une localisation.

Dans notre cas, nous utilisons GLiNER pour cette tâche d'identification d'entités nommées, en raison de son faible coût en ressources et de son efficacité démontrée.

Puisque les stupéfiants peuvent être facilement identifiés par des extractions simples *via* des expressions régulières avec une grande performance, il est inutile de chercher à détecter le type de stupéfiant: seule la quantité est recherchée. Puisque les quantités recherchées correspondent à des poids, on pourra demander alternativement de rechercher des quantités de stupéfiants ou des poids de stupéfiants. Les quantités ne correspondant pas aux unités de mesure souhaitées sont simplement supprimées, tandis que les poids, *a priori*, correspondent toujours à des unités de mesure souhaitées – c'est-à-dire, convertibles en grammes.

Pour éviter de devoir faire appel à un algorithme d'affectation linéaire (voir la section 2.2), nous n'indiquons pas au modèle de détecter l'ensemble des « quantités de stupéfiants ». À la place, nous tirons de l'extraction des types de stupéfiants par l'expression régulière une liste de stupéfiants, et nous accolons la quantité souhaitée au(x) type(s) de stupéfiants présent(s). Ainsi par exemple, si les expressions régulières identifient la cocaïne et l'héroïne parmi les stupéfiants présents dans le résumé de procédure, la liste d'entités sera [“quantité de cocaïne”, “quantité d'héroïne”], ou alternativement, [“poids de cocaïne”, “poids d'héroïne”].

Le cas où plusieurs quantités d'un même stupéfiant apparaîtrait dans les résumés de procédure, comme dans l'exemple de la section 2.2 page 15, n'est pas un problème puisque le modèle est en mesure de renvoyer plusieurs entités correspondantes si elles dépassent toutes un seuil défini du score de correspondance.

4 L'utilisation brute des modèles de langage

4.1 L'apport du SLM dans la recherche de modèles légers

Comme nous l'avons exposé avec UniversalNER (section 3.2), les grands modèles de langage (*Large Language Models* – LLM) sont en mesure de fournir des apports très intéressants dans la tâche d'identification d'entités nommées. Néanmoins, leur taille importante empêche généralement une utilisation directe – et même indirecte dans le cas d'UniversalNER – de ceux-ci si les capacités computationnelles sont limitées.

Pourtant, un LLM est en mesure d'extraire des entités arbitraires. Doté d'une grande flexibilité, seul le *prompt* d'entrée d'un modèle de langage conversationnel est à changer pour adapter la consigne à nos besoins. Bien évidemment, le *fine-tuning* du LLM permet d'augmenter ses performances dans la tâche souhaitée, mais procéder en *zero-shot* (sans *fine-tuning*) est tout à fait possible. Dans tous les cas, les grands modèles de langage sont difficiles à exploiter tels quels, d'autant plus que dans le cadre de données confidentielles comme les résumés des procédures de la police et de la gendarmerie, l'ensemble de ces modèles doivent être exécutés localement.

Dans ce cadre, les petits modèles de langage (*Small Language Models* – SLM) constituent une alternative intéressante. Ces modèles sont dits « petits » car ils se caractérisent par un nombre de paramètres nettement inférieur à celui des modèles linguistiques de grande taille. En effet, les grands modèles de langage comptent plusieurs milliards de paramètres, comme GPT-3 (175 milliards [17]) lancé en 2020 par OpenAI, ou Llama 3 (avec différentes versions de 8 à 405 milliards de paramètres) lancé en 2024 par Meta – et publiquement accessible. Les SLM, quant à eux, sont des versions de modèles de langage qui dépassent rarement le milliard de paramètres. Il s'agit souvent d'une version réduite ou « distillée » d'un plus grand modèle de langage. Par exemple, pour UniversalNER (section 3.2), c'est ainsi une version distillée de ChatGPT, réduite en paramètres, qui est utilisée. La distillation des SLM permet ainsi d'obtenir un modèle comportant beaucoup moins de paramètres mais d'une manière optimisée telle que les performances ne sont que légèrement réduites [18].

Ces modèles sont adaptés à un usage par un ordinateur sans GPU dédié, puisqu'un nombre moins important de calculs est nécessaire lors de l'inférence pour la génération du texte à renvoyer.

Dans notre cadre, un SLM en *open source*, téléchargeable et utilisable en local était nécessaire. Llama 3, depuis 2024, fournit une version de son modèle à 1 milliard de paramètres et multilingue. Il constitue ainsi une option intéressante pour notre cas d'usage.

4.2 La construction de l'instruction

La mission du SLM dans notre cas est donc de détecter le poids des différents stupéfiants détectés dans les résumés de procédure. Comme pour les NER, nous utilisons à notre avantage le fait que les stupéfiants ont déjà été extraits. Plutôt que de donner une instruction du type « Détecte les stupéfiants et leurs poids associés dans le texte suivant : », à partir du stupéfiant que l'on sait présent dans le résumé de procédure, par exemple la cocaïne, nous demandons plus précisément au SLM quelque chose du type : « Détecte le poids de cocaïne dans le texte suivant : ». Si plusieurs types de stupéfiants ont été extraits du résumé de procédure, alors il y aura autant d'instructions que de stupéfiants extraits, afin d'associer à chaque stupéfiant un poids. Cette approche se montre plus efficace que de demander au SLM le poids de tel stupéfiant mais aussi de tel autre stupéfiant dans une même instruction.

Pour que le modèle de langage comprenne bien la tâche qui lui est instruite, il est toujours utile de fournir au modèle d'autres éléments contextuels.

Tout d'abord il est possible de fournir un *system prompt*, la partie qui initialise le système, par exemple :

Tu es un assistant chargé d'extraire les poids exacts de différents stupéfiants dans des textes. Donne de façon synthétique les poids exacts associées aux types de stupéfiants présents dans le texte. Si la quantité est inconnue, indique Non connu.

Ensuite, il est bienvenu de donner des exemples fictifs ; il ne s'agit pas de *fine-tuning* mais cela permet de la même manière d'orienter le modèle vers la tâche souhaitée, en particulier si les instructions précédentes ne sont pas suffisamment claires ou sont involontairement ambiguës pour le modèle.

Il est important de couvrir l'ensemble des cas importants, en l'occurrence le cas où la quantité souhaitée est présente dans le résumé de procédure et le cas où elle y est absente. Les exemples sont ainsi construits :

Par exemple pour la proposition suivante : individu interpelle le 18/12 en train de vendre 2 grammes d'ecstasy et de l'héroïne. < Ecstasy : 2 grammes >

Pour la proposition suivante : individu interpelle le 18/12 en train de vendre 2 grammes d'ecstasy et de l'héroïne. < Héroïne : Non connu >

Enfin il faut inclure l'instruction finale : « Et pour la phrase suivante ? ». À cette instruction est accolée le résumé de procédure en question, suivi du stupéfiant dont on veut connaître la quantité, sur le modèle : « [Résumé de procédure]. < [Stupéfiant] : ». Il est ainsi demandé au modèle de compléter avec la quantité présente (ou la mention Non connu). Le modèle va logiquement terminer sa réponse par le signe >, nous permettant d'extraire facilement la quantité affichée.

Ainsi l'instruction globale utilisée suit systématiquement le schéma suivant :

- *System prompt*
- Exemples
- Texte à traiter

4.3 Calibrage de la température

Si cette méthode est simple, elle rencontre cependant quelques travers. L'utilisation de SLM, moins performants que les LLM, entraînent une compréhension réduite du sens des phrases, en particulier quand elles sont complexes. Il peut ainsi y avoir des contresens : parfois les quantités ne sont pas détectées alors qu'elles sont présentes ; parfois elles sont associées au mauvais stupéfiant ; enfin, pire encore, elles sont parfois complètement inventées alors que le chiffre en question n'apparaît jamais dans le résumé de procédure.

Ces « hallucinations » peuvent être partiellement supprimées par le réglage de la température. Il s'agit d'un paramètre qui permet d'augmenter ou de diminuer le degré de randomisation de la réponse. Moins il est élevé, plus la réponse est déterministe et ne risque pas de changer si on répète la même instruction. Cependant, même en réglant une température très faible, le modèle de langage préfère souvent inventer une quantité absente plutôt que de déclarer la quantité comme inconnue – il s'agit en réalité dans les deux cas pour le modèle de produire une réponse faite d'éléments absents du texte. En revanche, cela arrive rarement quand les résumés de procédure sont simples (notamment quand ils ne comportent qu'une seule phrase).

5 Performance des modèles et perspectives

5.1 Constitution d'un échantillon de test

Pour tester la performance de ces différentes méthodes, un petit échantillon de test a été tiré de l'ensemble des résumés de procédure dont la détection n'est pas permise par les bases Osiris, Objet ou les procès-verbaux électroniques, qui correspondent ainsi aux cas où l'utilisation des résumés de procédure est requise (voir la section 1.2).

Dans un premier temps, 250 résumés de procédure ont été tirés aléatoirement parmi cet ensemble. Sur ces 250 résumés de procédure, 42 ne contiennent pas de stupéfiant ; la détection de quantité associée y est donc caduque. Les 208 résumés de procédures restants contiennent 286 types de stupéfiants, il y a donc 286 quantités potentielles à extraire.

Les méthodes suivantes sont testées :

1. Expression régulière et affectation par l'algorithme de Jonker-Volgenant
2. GLiNER, avec détection de « quantités »
3. GLiNER, avec détection de « poids »
4. Llama-1B, avec une température à 0.1
5. Llama-1B, avec une température à 0.5
6. Llama-1B, avec une température à 0.9

Pour chaque méthode on calcule, sur la base des 286 types de stupéfiants, un indicateur de précision (le nombre de prédictions correctes sur l'ensemble des prédictions). La prédiction n'étant pas fondée sur un classificateur binaire, la notion de faux positifs, vrais négatifs, etc. n'aurait pas de sens ici. Néanmoins nous évoquons plus loin la possibilité d'utiliser d'autres mesures d'erreur qui ne sont pas dérivées de ces notions.

5.2 Performance des modèles

Globalement, les modèles GLiNER et Llama, en l'absence de fine-tuning, ont des performances inférieures à un algorithme déterministe robuste. Le modèle Llama possède la précision la plus faible, avec moins d'un stupéfiant sur deux détecté correctement ; le modèle invente régulièrement des quantités inexistantes, et ce même avec un réglage bas de la température : la précision s'établit à 0,40 pour une température à 0.9, 0,42 pour une température à 0.5, et 0.45 pour une température à 0.1. Le temps d'exécution est également long, puisque pour traiter les 250 résumés de procédure, plusieurs dizaines de minutes sont nécessaires.

Le modèle GLiNER performe globalement mieux, en partie puisqu'il n'invente aucune quantité qui serait absente du texte, par construction. Il extrait plus souvent l'information souhaitée quand il lui est demandé d'extraire des « quantités » (précision de 0.69) que des « poids » (précision de 0.67). De plus, il est plus rapide que le SLM puisqu'il faut moins de deux minutes pour traiter les 250 résumés de procédure. Tout ceci fait sens quand on sait que GLiNER peut être vu comme une version fine-tunée d'un modèle de langage spécifiquement pour la tâche d'extraction d'entités nommées.

Cependant, ces modèles en zero-shot performent pourtant moins bien que l'algorithme utilisant les expressions régulières, qui atteint une précision de 0.91 : 91 % des prédictions de quantités sont correctes (c'est-à-dire que, non seulement le poids est détecté dans le texte et correctement retranscrit, de manière exacte donc à la décimale près, mais il est aussi associé au bon stupéfiant ou est ignoré s'il ne correspond à aucun stupéfiant). Il est aussi très rapide, il suffit de quelques secondes pour obtenir les extractions sur les 250 résumés de procédure.

Modèle	Précision	Durée d'exécution par <i>manop</i>
Expression régulière + Jonker-Volgenant	0.91	< 0.1 s
GLiNER, avec détection de « quantités »	0.69	1 s
GLiNER, avec détection de « poids »	0.67	1 s
Llama-1B, avec une température à 0.1	0.45	30 s
Llama-1B, avec une température à 0.5	0.42	30 s
Llama-1B, avec une température à 0.9	0.40	30 s

TAB. 2 – Récapitulatif des performances

5.3 Regard critique sur l'indicateur de précision

En réalité, cet indicateur de précision ne suffit pas entièrement pour traduire la performance de l'algorithme. Les potentielles erreurs qui peuvent être effectuées en prédisant une quantité incorrecte ne sont pas toutes égales ; il y a en effet différents types d'erreurs qui sont assez communes mais dont l'impact est très différent.

Un premier type d'erreur commun pouvant se produire est l'omission d'une décimale : dans des cas comme « un sachet de 4 .5 grammes » comme on peut le lire parfois, plusieurs modèles sont susceptibles de l'interpréter comme 4 grammes, et la décimale, n'étant pas correctement attachée au reste du nombre, est ignorée par les modèles. Le modèle détecte alors 4 grammes au lieu de 4,5 grammes.

Un autre type d'erreur commun est l'erreur de conversion : dans un résumé de procédure mentionnant 500 kilogrammes, parfois et pour différentes raisons le modèle peut être amené à prédire 500 grammes. Une erreur d'importance similaire correspond au cas où une quantité très importante (par exemple, 1 tonne) est présente dans le texte mais non détectée par le modèle.

Une des utilisations possibles de l'extraction de ces poids est d'en tirer des poids moyens, ou des poids totaux sur un certain type de stupéfiant, éventuellement croisé à une autre variable. Dans cette optique, une erreur d'une décimale est moins préjudiciable qu'une erreur de conversion d'unité ou la non-détection d'une très grande quantité : dans ce dernier cas, il y a de grandes chances que cela ait un impact sur l'estimation du poids total ou du poids moyen.

Ainsi, les performances pourraient notamment être étudiées par le biais d'un indicateur de proximité. Plutôt que d'éliminer les prédictions incorrectes, on calcule l'erreur effectuée par la prédiction en prenant en compte l'écart relatif : en particulier, par le calcul de l'erreur quadratique moyenne, ou l'erreur absolue moyenne. Elles permettraient ainsi de différencier les écarts en fonction de leur amplitude.

6 Vers une approche hybride ?

Les trois approches prometteuses présentées jusqu'à présent (par les expressions régulières, par les modèles de NER généralistes et par les SLM) sont toutes susceptibles de fournir des résultats encourageants dans les cas simples, mais rencontrent des difficultés dans le cadre de résumés de procédures longs et à la syntaxe complexe.

Une dernière approche permet de pallier le problème des résumés de procédure complexes : il s'agit d'utiliser la capacité des SLM à résumer un texte.

Dans le cas d'un résumé de procédure « potentiellement complexe », une approche hybride en deux étapes peut être pertinente.

Tout d'abord, il est demandé au SLM de simplifier le résumé de procédure, en une ou deux phrases simples, en conservant impérativement l'ensemble des stupéfiants mentionnés avec leurs poids et leurs quantités. L'instruction peut prendre plusieurs formes ; après différents tests, la formulation qui est retenue est la suivante : « Tu es un assistant chargé de résumer des textes de sorte que l'information portant sur les quantités mentionnées soit facilement lisible, et associée à la substance en question. S'il n'y en a pas, indique 'Pas d'informations sur le sujet' ». Le terme « stupéfiant » ou « drogue » n'est volontairement pas mentionné : cette formulation permet de repérer des termes que le modèle n'aurait pas forcément détectés comme des stupéfiants, par exemple l'abréviation « rc » relative à la résine de cannabis. Cela diminue la probabilité d'observer que le modèle insère des stupéfiants dans les résumés de procédure qui n'en mentionnent pas.

À cette instruction, est ajoutée un exemple de simplification de résumé de procédure qui montre une manière de résumer simplement les informations relatives aux quantités de stupéfiants.

Ensuite, sur le résumé de procédure simplifié, les méthodes précédemment présentées ont davantage de chances de détecter les poids correctement que sur les résumés de procédures complexes dont les tournures rendent parfois l'extraction difficile.

Ainsi, il est possible d'extraire les stupéfiants par les expressions régulières : c'est la même expression régulière que celle présentée en section 2.1 qui est utilisée, puisque la reformulation du résumé de procédure est censée fournir un résultat plus simple que le résumé de procédure initial, mais sans différence rédactionnelle par rapport à un résumé de procédure simple qui aurait pu être écrit par les services de sécurité intérieure.

Cependant il est tout à fait possible d'utiliser plutôt les modèles de NER ou les SLM sur les résumés de procédure simplifiés.

Il s'agit d'un compromis intéressant dans le cadre de capacités computationnelles limitées ne permettant pas l'utilisation de modèles de langage très lourds – et donc potentiellement, très puissants –, ou de modèles d'extraction d'entités nommées très fins.

Les premiers résultats observés montrent que les SLM peuvent être de très bons outils pour cette tâche de simplification, où ils performant davantage que pour extraire l'information sur les types de stupéfiants. Le résumé de procédure simplifié est un objet beaucoup plus facile à gérer pour l'utilisation d'expressions régulières, le modèle GLiNER ou l'utilisation successive du même SLM pour la tâche d'extraction de quantités.

En revanche, l'ajout d'une étape supplémentaire augmente le risque de perte d'information importante dans le processus : des stupéfiants peuvent être omis par le SLM lors de la simplification du résumé de procédure, rendant toute détection ultérieure de quantité impossible. Le contrôle des hallucinations du modèle est donc primordial.

Conclusion

Dans le cas idéal, l'extraction de quantités dans un texte peut passer par l'utilisation de LLM puissants ou des modèles spécialisés dans l'extraction d'entités nommées (NER) plus directement encore. La méthode classique revient à fine-tuner sur nos données ces modèles d'analyse textuelle sur notre tâche d'extraction des quantités ; une fois entraîné, le modèle atteint des performances supérieures à un modèle utilisé en *zero-shot* – c'est-à-dire sans entraînement. Dans notre cas, les textes utilisés pour l'entraînement correspondraient à des procédures dont les quantités sont déjà connues par le biais d'autres bases, et n'auraient donc aucune différence rédactionnelle notable avec les textes à prédire, puisqu'ils sont rédigés par les mêmes services de sécurité intérieure.

Dans le cas où l'on cherche des méthodes légères, adaptées à un environnement computationnel plus restreint, l'utilisation de ces modèles devient rapidement irréaliste.

Les avancées en la matière fournissent des alternatives intéressantes : les petits modèles de langage (SLM) disposent d'un nombre de paramètres bien moins élevé que les LLM, permettant leur utilisation sur la plupart des machines ; des modèles de NER existent en version légère en étant déjà fine-tunés sur la détection d'un nombre d'entités possibles important, voire illimité, dans le cas de GLiNER, ou d'UniversalNER bien que ce dernier soit plus volumineux.

Cependant, ces modèles bien qu'optimisés pour une réduction minimale des performances, ne rivalisent pas pour ce genre de tâches avec les modèles les plus fins et les plus performants ; c'est en réalité inhérent à la tâche demandée : l'extraction de quantités et leur affectation à chaque type de stupéfiant présent dans le texte est une tâche complexe qui demande au modèle une certaine « compréhension » de la logique inhérente de la phrase, sans laquelle le modèle est rapidement amené à faire des contresens où à détecter des stupéfiants là où il n'y en a pourtant pas.

Dans ce genre de cas, l'utilisation directe ou indirecte des modèles de langage n'est pas toujours à recommander, quand il existe des alternatives déterministes plus fiables, bien qu'elles soient également moins flexibles ; c'est le cas ici avec l'algorithme qui combine l'utilisation d'expressions régulières et l'algorithme de Jonker-Volgenant pour l'affectation des quantités aux stupéfiants.

Cependant, l'utilisation de modèles de langage, même les plus simples, fournit une aide importante pour des sous-tâches de l'extraction de quantités : ainsi, résumer ou simplifier les textes par le biais de modèles de langage en amont de l'extraction peut constituer une option intéressante pour augmenter les performances de l'algorithme déterministe d'extraction.

Bibliographie

1. SSMSI. *Insécurité et délinquance en 2024 : bilan statistique et atlas départemental* 144–149. <https://www.interieur.gouv.fr/content/download/138329/1093108/file/SSMSI-BA2024.pdf> (2025).
2. ONUDC. *Terminologie et informations relatives aux drogues*. <https://www.un-ilibrary.org/content/books/9789210045445> (United Nations, 2019).
3. Korf, D., Benschop, A. & Wouters, M. Differential responses to cannabis potency: a typology of users based on self-reported consumption behaviour. *International Journal of Drug Policy* **18**, 168–176 (2007).
4. Vrolijk, R. Q. *et al.* Size matters: comparing the MDMA content and weight of ecstasy tablets submitted to European drug checking services in 2012–2021. *Drugs, Habits and Social Policy* **23**, 207–219 (2022).
5. Akgül, M. The linear assignment problem. *Combinatorial Optimization: New Frontiers in Theory and Practice*, 85–122 (1992).
6. Jonker, R. & Volgenant, T. Improving the Hungarian assignment algorithm. *Operations research letters* **5(4)**, 171–175 (1986).
7. Nadeau, D. & Sekine, S. A survey of named entity recognition and classification. *Linguisticae Investigationes* **30**, 3–26 (2007).
8. Li, J., Sun, A., Han, J. & Li, C. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering* **34**, 50–70 (2020).
9. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
10. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019).
11. Martin, L. *et al.* CamemBERT: a Tasty French Language Model. *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics* (2020).
12. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. & Androutsopoulos, I. LEGALBERT: The Muppets straight out of Law School. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904 (2020).
13. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2019).
14. Touchent, R., Romary, L. & De La Clergerie, E. CamemBERT-bio: Un modèle de langue français savoureux et meilleur pour la santé. *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1: travaux de recherche originaux–articles longs*, 323–334 (2023).
15. Zhou, W., Zhang, S., Gu, Y., Chen, M. & Poon, H. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *The Twelfth International Conference on Learning Representations* (2024).
16. Zaratiana, U., Tomeh, N., Holat, P. & Charnois, T. Gliner: Generalist model for named entity recognition using bidirectional transformer (2023).
17. Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).
18. Schick, T. & Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2339–2352 (2021).

Annexe

Codage des types de stupéfiants inspiré de la nomenclature de l'ONU DC

En italique sont indiqués les stupéfiants absents de la nomenclature de l'ONU DC.

1. Cannabis	1A. Résine de cannabis 1B. Herbe de cannabis 1C. Huile de cannabis 1D. <i>Pied de cannabis</i> 1E. <i>Pollen de cannabis</i> 1F. <i>Graines de cannabis</i> 1Z. Cannabis non caractérisé
2. Cannabinoïdes de synthèse	2A. Cannabinoïdes de synthèse
3. Opium et opiacés	3A. Produits de l'opium 3B. Morphine 3C. Codéïne 3D. Thébaïne 3E. Héroïne 3Z. Opiacés non caractérisés
4. Opioïdes	4A. Fentanyl (et assimilés) 4B. Méthadone 4C. AH-7921 4D. <i>Nitazènes</i> 4E. <i>Tapentadol</i> 4F. <i>Oxycodone</i> 4Z. Opioïdes non caractérisés
5. Coca et cocaïne	5A. Produits du cocaïer 5B. Cocaïne simple 5C. Crack 5D. Freebase 5E. Ecgonine
6. Stimulants de type amphétamine	6A. Amphétamine 6B. Méthamphétamine 6C. Ecstasy/MDMA et assimilés 6D. Cathinones synthétiques 6E. <i>Feuilles de khat</i> 6F. Autres stimulants synthétiques du SNC ^a
7. Dépresseurs du SNC	7A. GHB 7B. GBL
8. Hallucinogènes	8A. LSD 8B. Tryptamines 8C. Cactus peyotl, mescaline 8D. Champignons psilocybes 8E. Hallucinogènes de synthèse 8F. PCP 8G. <i>Kétamine</i>

^aSNC : Système Nerveux Central.