Présentez votre texte chunké à la main en justifiant de certains choix le cas échéant

Le texte que j'ai analysé s'intitule *« Pour freiner la fonte des glaces, le gouvernement fait installer 3000 climatiseurs sur la banquise* ». Présentée comme une solution *"écologique et économique",* cette idée inclut aussi des propositions comme construire un grand réfrigérateur pour la banquise ou collecter la glace des frigos des Français. C'est une satire dénonçant l'inaction politique réelle face à l'urgence climatique.

Le texte analysé contient 254 chunks répartis en 23 types ou catégories différents. Parmi les plus fréquents se trouve le groupe nominal noté **N**. Il se compose généralement d'un déterminant suivi d'un nom. Ces déterminants peuvent être définis « *le gouvernement* », indéfinis « *un porte-parole* », possessifs « *notre idée* », démonstratifs « *ce genre* » ou quantificateurs « *tous les cent kilomètres carrés* ». Le groupe peut aussi inclure des adjectifs épithètes ou attributs qui ne prennent pleinement sens qu'à l'intérieur du chunk, avec « *des programmes cohérents* » ou « *un très gros réfrigérateur* ». Concernant les noms isolés (non précédés de déterminant) comme « *Reportage* », « *totalité* » ou « *efficacité* », ils n'ont pas été classés comme N pour éviter d'intégrer des cas atypiques dans le lexique. Autre type fréquent : les groupes prépositionnels nominaux **PN**. Ils sont constitués d'une préposition (simple ou contractée) puis d'un groupe nominal. Leur fonction principale est d'agir en tant que compléments circonstanciels (de lieu, de moyen, de cause, d'agent..) ou compléments du nom (« *sur la banquise* », « *avec le WWF* », « *par un immense pont aérien* »). Les prépositions peuvent être simples (de, à, dans), contractées (du, au), ou composées avec un déterminant démonstratif, possessif, quantitatif ou numéral. Elles introduisent une relation syntaxique et appellent un complément. Les marqueurs observés incluent P (*de*), P + Det (*sur* + *la*), Pdet (*du*), ou P + ProRel (*dans lequel*). Dans l'ensemble, ces groupes n'ont pas posé de difficulté particulière lors du processus de chunking manuel. Les adjectifs ou adverbes modifiant le nom ou précisant le complément sont généralement intégrés au chunk afin de préserver la cohérence syntaxique de la phrase.

Dans mon corpus, les éléments de ponctuation (guillemets et signes de ponctuation) ont été chunkés séparément pour préserver la clarté discursive. **GO** et **GF** (guillemets ouvrants «, et fermants ») sont séparés car ils signalent des transitions discursives. **Pctnf** (ponctuation non finale ,) sépare des unités internes à la phrase et **Pctf** (ponctuation finale .) marque la clôture des phrases. Les **BR** sont réservés pour les retours à la ligne pour les changements de paragraphe. Eux aussi ne sont pas intégrés aux chunks syntaxiques, marquant la structuration textuelle. Tous ces marqueurs se retrouvent intégrés dans le lexique pour les identifier automatiquement. Les conjonctions de coordination comme *(et, ou, mais, donc)* sont catégorisées sous **Coord** et annotées CC dans le lexique. Elles servent à relier des unités syntaxiques et ne sont pas incluses dans les chunks.

Pour ce qui est des adverbes **Adv**, j'ai choisi de les chunker séparément quand c'était possible. Ils disposent parfois d'une autonomie et peuvent être déplacés ou isolés sans modifier la structure de la phrase. Les intégrer systématiquement dans d'autres chunks alourdirait les règles de segmentation et risquerait de produire du bruit. Il y a à la fois des adverbes simples

comme « ensuite », « alors », « mieux » et des locutions adverbiales composées comme « de plus » ou « voire même ». Certains apparaissent au sein de chunks comme les compléments subordonnés (CS dans des Csub, « *peut-être* que », « *voire même* que ») ou en combinaison avec des adjectifs (*surtout* écologique, *trop* chaud) ou des Pver (pour agir *efficacement*). Dans ces cas-là, ils ont été intégrés dans des chunks plus larges. Concernant les adjectifs **Adj**, ils sont chunkés à part uniquement lorsque cela était pertinent (comme avec les adverbes). Par exemple, dans des chunks comme « surtout *écologique* » ou « trop *chaud* », l'adjectif est conservé dans un chunk spécifique lorsqu'il est accompagné d'un modifieur. Dans l'extrait « des solutions *logiques* et *simples* », j'ai choisi de chunker « *logiques* » et « *simples* » séparément. Ces deux adjectifs sont coordonnés par « et » et modifient le même nom « solutions » de manière indépendante. Les séparer permet de respecter la structure syntaxique de la coordination. Je n'ai pas constitué de lexique pour les adjectifs car leur usage est trop contextuel. De nombreux adjectifs sont intégrés directement dans des groupes nominaux (PN) où ils ont une fonction descriptive (ex. : « par un immense pont *aérien*», dans une posture *politique* »). Isoler systématiquement les adjectifs aurait fragilisé la structure syntaxique et rendu le chunker moins fiable.

Les conjonctions subordonnées formant des chunks de type **Csub** introduisent des propositions subordonnées. Dans le lexique, elles sont toutes annotées avec le marqueur CS, ce qui permet de les reconnaître lors du traitement. Les conjonctions simples « *qu', que, quand, lorsque* » sont marquées **CS** dans le lexique et les conjonctions composées avec adverbes : *« voire même qu', peut-être que »* sont aussi chunkées [Csub] mais issues de la combinaison Adv + CS. Dans « *Face à* », le mot « *face* » est normalement un nom commun, tandis que « *à* » est une préposition. Lorsqu'ils apparaissent ensemble, ils forment une locution prépositionnelle **LP** qui introduit un groupe prépositionnel (ex. : « *face à l'urgence* »). Il n'était donc pas possible de segmenter « *face* » et « *à* » séparément car cela aurait rompu la structure syntaxique et généré du bruit. « Face à » a été chunké comme un groupe figé même si seul « *à* » appartient à la catégorie des prépositions dans le lexique.

Le pronom démonstratif « *celle* » a intégré le lexique sous la catégorie **Pdem**. En tant que pronom, il possède une autonomie remplaçant un groupe nominal sans nécessiter d'éléments associés. Il reprend un référent (« *l'idée de larguer 3000 climatiseurs* ») et fonctionne comme un noyau nominal autonome. Dans la phrase « *et vous ?* », le pronom personnel sujet **«** *vous* **»** est isolé en fin de séquence pour interpeller directement le lecteur. Il est identifié dans le lexique sous la catégorie **PPS** (Pronom Personnel Sujet) et chunké séparément. Le pronom relatif sujet **«** *qui* **»** a été traité comme un **PRelSuj**. Il est identifié dans le lexique sous ce marquage pour signaler le début d'une subordonnée relative. « *qui* » relie la subordonnée à l'antécédent tout en occupant la fonction de sujet du verbe de cette subordonnée (« ce sont plus de 3 000 climatiseurs *qui* vont être largués »). Chunker « qui » de manière autonome permet de préserver les relations entre propositions principales et subordonnées relatives. Le pronom relatif « *où* » a été identifié dans le lexique sous la catégorie **PRel**. Il est chunké séparément pour signaler le début d'une subordonnée relative circonstancielle. Dans l'exemple « soutenir des programmes cohérents et *où* les solutions ne viennent pas d'un manuel », « *où* » introduit une subordonnée.

Le groupe verbal « *en annonçant* » a été identifié comme un **PVant** (Préposition + Participe Présent). Il est chunké séparément car il forme une unité syntaxique autonome, exprimant la manière dont l'action principale est réalisée (« agir vite et bien »). Les séquences composées d'une préposition suivie d'un verbe à l'infinitif ont été traitées comme des chunks de type **PVer**. Ils sont reconnus par deux marqueurs principaux : la présence d'une préposition (pour, à, de..) et suivie d'un verbe à l'infinitif marqué par la terminaison -er, -ir, ou -re. Cela permet de regrouper des constructions syntaxiques exprimant des buts ou des moyens comme dans « pour freiner », « à proposer » ou « de larguer ». Dans mon corpus, j'ai isolé les groupes Sujet + Verbe **SV** en chunkant ensemble le sujet, son verbe conjugué, ainsi que d'éventuels modificateurs ou particules négatives. Les marqueurs utilisés (Pro, PPS, ProS, Mod, Neg, Pref) permettent de reconnaître la structure sujet-verbe incluant les cas de pronominalisation, de négation, de modification adverbiale ou de pronom réfléchi. L'ensemble « *a-t-il ajouté* » a été chunké comme **SVInv**, car il présente une inversion du sujet et du verbe. La détection repose sur la présence du pronom inversé (-il) annoté dans le lexique sous PPS.

Les verbes conjugués ont été isolés dans des chunks de type **V** en s'appuyant sur des marqueurs morphologiques simples : terminaisons caractéristiques (-er, -é), auxiliaires (a, sont, est), particules négatives (ne, pas). Cela permet de chunker efficacement la majorité des structures verbales du corpus. Certains cas comme « précise » soulèvent une limite : n'étant pas couvert par les marqueurs classiques, il a été chunké que grâce à une connaissance contextuelle du corpus. Dans les phrases où des verbes comme « fait installer », « est », « sera », « a cependant rappelé » apparaissent, le groupe qui précède n'est pas un pronom personnel sujet (PPS) mais un groupe nominal (N) avec « le gouvernement », « l'objectif », « la dernière option ». Dans mon système de chunking, la catégorie **SV** est réservée aux structures où un pronom personnel sujet est directement suivi de son verbe conjugué (« nous agissons », « il est temps »). Quand le sujet est un groupe nominal, il est chunké séparément en N, et le verbe est chunké seul en V.

Le verbe « affirmant » est chunké **Vant** : il s'agit d'une forme verbale au participe présent avec la terminaison en -ant. Il a été chunké séparément car il exprime une action complémentaire. Les formes verbales *« réfléchir, tripler, construire, installer »* ont été chunkées individuellement en **Ver**. Elles sont identifiées grâce à leurs marqueurs morphologiques VInf -er, -ir, -re, pour les infinitifs. Ces verbes apparaissent sans être directement précédés par un pronom personnel sujet, mais soit après des coordinations (*ou*), soit introduits par un verbe principal (*a annoncé*), soit rattachés à des groupes nominaux.

Présentez la base de règles et le lexique conçus à partir de ce texte et leurs principes (mécanisme des règles, catégorisations du lexique)

L'ordre des règles dans le fichier regles.txt a été structuré en suivant une logique allant du plus spécifique et complexe avec plusieurs combinaisons au plus général. Plusieurs versions concernant l'ordre des règles ont été testées afin d'observer l'effet de chaque règle sur l'analyse globale. Certaines règles placées trop tôt produisaient des analyses incorrectes ou entraient en

conflit avec des structures plus précises. Les règles les plus structurées comme celles pour les groupes sujet-verbe (PPS + Mod, PPS + Neg + _) ou les constructions inversées (Mod + -PPS + _) doivent être en premier car elles reposent sur des éléments identifiables et peu ambigus. À l'inverse, les règles plus larges comme *er => [VInf ou P + _ => [PN doivent être placées plus bas dans la hiérarchie pour ne pas interférer.

Chaque entrée du lexique est structurée sur deux colonnes séparées par une tabulation : d'un côté la forme exacte du mot telle qu'elle apparaît dans le texte, et de l'autre son étiquette grammaticale (par exemple ADV, Det, CS). Cela permet de distinguer la chaîne de caractères à reconnaître de la catégorie à attribuer. Le contenu du lexique est composé de mots grammaticaux invariables ou très fréquents, tels que les adverbes (*bien, peut-être, plutôt*), les conjonctions (*et, mais, que, lorsque*), les déterminants (*le, la, des, notre*), les prépositions (*de, à, en, sur*), les pronoms personnels (*il, elle, nous*), les pronoms relatifs (*qui, où*), les auxiliaires (*est, sont, avons*), les marqueurs de négation (*ne, pas*), ainsi que les formes contractées (*du, au*) et les signes de ponctuation (*.*, *?*, *«*, *»*).

La règle ***ique => [Adj]** (et sa variante plurielle *iques → [Adj]) repose sur un principe de catégorisation basé sur les suffixes. En français, de nombreux adjectifs se terminent par -ique (ex. écologique, logique, politique). Mais cette règle ne peut pas s'appliquer aux autres expressions car elles ne respectent pas cette terminaison. Par exemple, *trop chaud* et *aussi cool* sont des groupes adjectivaux composés d'un adverbe (*trop*, *aussi*) suivi d'un adjectif (*chaud*, *cool*). Sans une règle du type ADV + Adj => [Adj], ces séquences ne peuvent pas être identifiées comme adjectivales. Inclure des adjectifs spécifiques comme *simple*, *chaud*, ou *cool* dans le lexique figerait l'analyse et la rendrait moins généralisable. Ce serait une solution ponctuelle mais inefficace à grande échelle. Chaque adverbe de la liste (*bien, plus, surtout, voire même, peut-être, ensuite, alors, mieux, plutôt*) est reconnu grâce à une combinaison de lexique et de règles de type **ADV => [ADV.** Ces adverbes pour la plupart sont figés et fréquents. Leur identification repose sur leur présence dans le lexique, car ils ne sont ni dérivables (même terminaison), ni prévisibles dans la phrase. Pour les adverbes composés ou locutions adverbiales *voire même* ou *peut-être*, leur reconnaissance repose sur leur traitement comme unités lexicales entières. Certaines formes (*de plus*) sont reconnues via la règle **P + ADV => [ADV**, rendue possible par le fait que toutes les prépositions (P) sont présentes dans le lexique.

Le lexique associé aux conjonctions de coordination se compose de *et*, *ou*, *mais* et *donc*. Ces mots étiquetés CC (conjonctions de coordination) sont reconnus par leur présence dans le lexique, composé de mots invariables. La règle appliquée **CC -> [Coord** consiste en un simple transfert de la catégorie lexicale vers une catégorie syntaxique. Pour les conjonctions de subordination *qu'*, *que*, *quand*, *lorsque*, étiquetées CS, leur reconnaissance repose sur la règle **CS => [CSUB.** Cette règle permet de projeter directement les conjonctions identifiées vers une catégorie syntaxique [CSUB] pour leur rôle d'introducteurs de subordonnées. Ces conjonctions sont des mots figés, ce qui justifie leur inscription dans le lexique. La base tient compte de combinaisons entre adverbes et conjonctions *voire même qu'* ou *peut-être que*, qui sont analysées à l'aide de la règle **Adv + CS => [CSUB.**

Trois types de ponctuation sont distingués : la ponctuation finale (Pctf), la ponctuation non-finale (Pctnf) et les guillemets ouvrants et fermants (GO et GF). Les signes comme le point (.) et le point d'interrogation (?) sont étiquetés Pctf car ils marquent la fin d'une phrase et sont projetés à l'aide de la règle **Pctf => [PCTF.** La virgule quant à elle, est classée comme Pctnf puisqu'elle signale une séparation sans rompre la structure syntaxique ; elle est traitée via la règle **Pctnf => [PCTNF.** Les guillemets ouvrants («) et fermants (») sont étiquetés respectivement GO et GF et associés à leurs règles dédiées (**GO => [GO, GF => [GF),** servant à baliser le discours rapporté. Le lexique utilisé ici est entièrement figé.

Pour les groupes nominaux (N), le lexique regroupe les déterminants (Det) définis (*le, la, les, l'*), indéfinis (*un, une, des*), possessifs (*nos, leur, notre*), démonstratifs (*ces, ce*) et les expressions numériques comme *3 000*, identifiées par l'expression régulière Num:[0-9]+. Une des règles associées est : **Det + _ => [N** où tout mot ou groupe de mots précédé d'un déterminant est interprété comme un groupe nominal ([N]). D'autres règles viennent ensuite élargir cette reconnaissance pour couvrir des N plus long comme **Det + _ + _ => [N**. La règle **Det + Det + _ => [N** permet d'identifier des groupes nominaux dans lesquels un numéral est combiné avec un déterminant comme dans « *les 3 000 climatiseurs »*. La règle **[0-9]+ + _ => [N** s'applique spécifiquement aux structures suivies d'un nom avec « *3 000 climatiseurs »*. En revanche, la règle Det + _ + _ + _ => [N qui aurait permis de reconnaître des groupes nominaux comportant trois mots après le déterminant comme « *des très grosses pyramides »* n'a pas été retenue. Même si elle permettrait de capter des syntagmes nominaux complexes, sa formulation pourrait conduire le système à interpréter à tort comme groupes nominaux des séquences non nominales (verbes, prépositions, conjonctions ou autres éléments). Le mot « celle » en tant que pronom démonstratif, est étiqueté Pdem. Le lexique contient cette entrée et la règle appliquée suivante **Pdem => [PDEM,** projette cette forme lexicale vers une catégorie syntaxique.

Pour les groupes prépositionnels nominaux (PN), le lexique comprend des prépositions (*à, de, sur, en, dans, par, pour, avec, contre*) sous leurs formes simples P (*de, à*) et contractées PDet (*du, au*). Dans les règles syntaxiques associées, on trouve la règle **P + Det + _ => [PN** utilisée pour « *sur la banquise »* ou « *avec le WWF »*. La règle **P + _ => [PN** permet quant à elle d'englober par exemple « *en partie »* ou « *d'élétroménager »*. D'autres règles s'appliquent à des cas spécifiques, comme les formes contractées via **PDet => [PN.** Dans des expressions comme « *de grande collecte »* ou « *de trois mille climatiseurs »*, la règle **P + _ + _ => [PN** permet de capturer des PN plus longs qui ne seraient pas couverts par une règle plus simple de type P + _ -> [PN. Le pronom personnel sujet *« vous »* est reconnu grâce à sa présence dans le lexique (PPS : elle, nous, vous) et permet la règle associée est **PPS => [PPS** pour « *et vous ? ».* Aussi, le pronom relatif sujet **«** *qui »* étiqueté PRelSuj a une règle simple et directe **PRelSuj => [PRelSuj.** « *Où* » en tant que pronom relatif étiqueté PRel est identifié dans le lexique avec la règle **PRel => [PREL.**

Avec la forme verbale *« en annonçant »* comme groupe prépositionnel verbal au participe présent, étiqueté PVant, la préposition « *en »* est présente dans le lexique sous P. La règle utilisée ici est **P + *ant => [PVant**. Lorsqu'une préposition est suivie d'un mot se terminant par -ant (typique du participe présent), l'ensemble est projeté en un groupe PVant.

Cette règle permet de reconnaître des formes comme (*en annonçant*, *en expliquant*) en les détectant grâce à leur suffixe. Pour les groupes prépositionnels verbaux à l'infinitif, étiquetés PVer, comme (*pour freiner*, *à proposer*, *de rejeter)*, la structure est composée d'une préposition (P) suivie d'un verbe à l'infinitif. Le lexique mobilisé dans cette section comprend les prépositions : *pour, de, à, par, en*. Les règles syntaxiques s'appuient sur la terminaison du verbe **: P + *er => [PVer** pour les verbes du premier groupe (*freiner*, *proposer*) et **P + *ir => [PVer** pour ceux du deuxième groupe (*agir*, *refroidir*). Ces terminaisons en -er et -ir sont suffisamment régulières pour permettre une reconnaissance fiable, contrairement à celles en -re plus variés morphologiquement. En revanche, la règle P + Pro + Inf -> [PVer] qui permettrait d'analyser *de les rejeter* n'a pas été retenue, mais plutôt **P + _ + *er => [PVer**. L'élément *les* peut être ambigu : c'est un pronom objet mais aussi un déterminant. Cela aurait introduit un risque de surgénéralisation. **P + *ir + _ => [PVer** permet de tenir compte des adverbes qui ne sont pas chunkés séparément.

Le lexique mobilisé pour les SV inclut les pronoms personnels sujets (PPS) comme (*elle*, *nous*, *vous)*, les pronoms impersonnels (ProS) tels que (*il*, *on*, *c')*, et les pronoms indéfinis (Pro) comme (*ce)*. Il comprend aussi des particules pronominales (Pref) (*se*, *s'*) pour les verbes pronominaux, ainsi que des verbes auxiliaires ou modaux (Mod) comme (*est, sont, avons, sommes)*. La règle **PPS + Mod + _ => [SV** couvre les formes comme « *nous avons mobilisé »***, PPS + _ => [SV** s'applique aux cas comme « *nous devons »***; Pro + Mod + ADV => [SV** permet de reconnaître des constructions avec « *ce sont plus »*. Des règles plus spécifiques gèrent les formes pronominales **PPS + Pref +** _ comme « *il s'agissait »*, ainsi que les structures négatives **PPS + Neg + _ => [SV** « *elle ne fonde »*. Les structures via **ProS + _ => [SV** permet d'identifier « *il fait »,* et « *c'est tout »* est capturée par **ProS + Mod + _ => [SV**. La structure inversée sujet-verbe avec « *a-t-il ajouté »*, classée sous le type SVInv, correspond à une inversion interrogative où le pronom sujet suit le verbe accompagné d'un trait d'union et d'un "t". La règle associée **Mod + *-PPS + _ => [SVInv** permet de projeter cette séquence comme un groupe sujet-verbe inversé où le * en préfixe de la règle indique qu'il prend un auxiliaire suivi du pronom inversé.

Les groupes verbaux (V) s'appuient sur le lexique des verbes modaux et auxiliaires tels que (*a, est, sont, avons, avez, sommes)* étiquetés Mod, les infinitifs comme (*être)* (ModInf), ainsi que les particules de négation Neg comme (*ne, pas, n')*. Il inclut également certains adverbes fréquents (*bien, plus, surtout, ensuite, plutôt*). Les règles permettent de reconnaître des formes variées : des infinitifs (**_ + *er => [V** comme dans « *fait installer »*), des temps composés (**Mod + *é => [V** pour « *a décidé*, *a martelé »*), ou encore des passifs (**ModInf + _ => [V** pour « *vont être largués »*). D'autres règles prennent en charge les formes conjuguées simples comme « *est, sont autant »* (avec **Mod => [V** et **Mod + _ => [V** ainsi que les structures avec négation par *ne...pas* (**Neg + _ + Neg => [V**). J'ai choisi de ne pas introduire une règle du type *ent -> [V] car cette terminaison peut également apparaître dans des noms (*événement*, *fondement*) des adjectifs (*différent*, *transparent*) ou des adverbes (*évidemment*), engendrant de l'ambiguïté. L'ajout de règles comme ***er => [VInf** et ***ir => [VInf** permet d'identifier des infinitifs isolés qui ne sont pas précédés de prépositions ou d'auxiliaires. En revanche, les règles similaires pour les formes en -re ou -ant sont bien plus sujettes à ambiguïté (comme *sourire* ou *manifestant*).

Présentez les principes fonctionnels du moteur à base de règles

Le script *chunker.py* commence par importer la librairie **re** qui permet d'utiliser des expressions régulières et aussi **ElementTree** qui permet de générer et manipuler des fichiers XML. **datatime** permet d'accéder à la date du jour dans le fichier XML pour indiquer quand l'analyse a été effectuée. Le script lit un texte brut en français, l'analyse en chunks, et génère deux sorties : un fichier XML structuré et une page HTML pour la visualisation.

La fonction **def charger_lexique(nom_fichier)** lit le fichier de lexique (lexique.txt). Chaque ligne du fichier contient un mot et sa catégorie séparés par une tabulation. Tous les mots sont convertis en minuscules pour éviter les erreurs dues à la casse. Le lexique permet d'associer à chaque mot une catégorie grammaticale (déterminant, adjectif, pronom…). **def charger_regles(nom_fichier)** lit le fichier des règles (regles.txt). Chaque ligne contient une partie gauche où les conditions de déclenchement sont séparées par +, et une partie droite avec la catégorie correspondante. Les commentaires (//) sont ignorés. Le tout est stocké dans une liste de tuples où chaque tuple contient une liste de conditions et la catégorie de sortie. Ces règles sont utilisées pour identifier automatiquement des chunks comme les groupes nominaux (N), verbaux (V), ou prépositionnels (PN), en analysant les suites de catégories dans le texte Pour la fonction **def pretokeniser_texte(texte_brut),** elle prépare le texte brut avant tokenisation. Elle ajoute des espaces autour de la ponctuation, gère les contractions avec apostrophe (l', d') pour les séparer proprement et normalise les espaces. Ce prétraitement facilite la tokenisation automatique en s'assurant que la ponctuation et les contractions ne perturbent pas l'analyse. **def charger_texte(nom_fichier)** lit le texte depuis un fichier (texte.txt). Elle remplace les sauts de ligne par le marqueur BR pour signaler des paragraphes. Elle renvoie une liste de mots (tokens) en supprimant les espaces en trop.

La fonction essentielle **def regle_satisfait(tokens, lexique, position, conditions)** vérifie si une séquence de mots respecte les conditions d'une règle à partir d'une position donnée. Chaque condition correspond à un élément de la partie gauche d'une règle (Det, ADV, *er, [0-9]+..). La fonction compare les mots du texte aux conditions suivantes : _ accepte un mot inconnu du lexique (non catégorisé), *er, *ir, *ant, accepte un mot se terminant par ce suffixe (souvent un verbe), [0-9]+ accepte un mot composé uniquement de chiffres (comme un nombre), *-PPS accepte une forme inversée avec pronom sujet suffixé comme dans « a-t-il », une catégorie lexicale (Det, ADV) vérifie que le mot a bien cette étiquette dans le lexique. Si tous les mots alignés à partir de la position donnée satisfont les conditions de la règle, la fonction retourne True et la liste des mots correspondants. Sinon, elle renvoie False et une liste vide. Cette fonction est utilisée par le script pour appliquer les règles de segmentation à chaque position du texte tokenisé dans le but d'identifier des chunks.

La fonction **appliquer_regles(tokens, lexique, regles)** applique les règles de segmentation sur une séquence de mots (tokens). Elle parcourt le texte mot par mot et à chaque position, teste toutes les règles dans l'ordre. Si une règle est satisfaite, elle extrait les mots concernés, les regroupe dans un chunk avec l'étiquette définie par la règle, et avance le pointeur de lecture

d'autant de mots que nécessaires. Les autres règles ne sont alors pas testées pour cette position (grâce à un break) donnant la priorité à la première règle applicable. Si aucune règle ne s'applique à une position donnée, le mot est simplement conservé tel quel sans étiquette (non chunké). Le résultat final est une liste dans laquelle chaque élément est soit : un triplet (catégorie, contenu, règle utilisée) pour les mots regroupés en chunk, soit un mot seul (non chunké) si aucune règle ne s'est appliquée. L'ordre des règles dans la liste est crucial : les règles les plus spécifiques doivent précéder les plus générales pour éviter des erreurs de segmentation. Le pointeur n'effectue aucun retour en arrière.

Pour générer un fichier XML à partir de la liste de chunks obtenue, la fonction **generer_xml_chunked(chunks,fichier_xml)** construit un document structuré hiérarchiquement. L'élément racine <texte> contient deux attributs : date (la date du jour) et src (la source du texte). Chaque paragraphe est représenté par une balise <paragraphe> contenant une ou plusieurs balises <phrase>. Chaque chunk est encodé sous forme d'un élément <chunk> avec deux attributs : cat pour la catégorie grammaticale (ex. : N, PN, SV...) et regle pour la règle appliquée. Les mots non analysés sont marqués avec cat="UNK" et regle="aucune". Les retours à la ligne (BR) déclenchent la création d'un nouveau paragraphe et les points (.) ouvrent une nouvelle phrase. Le fichier est enregistré au format XML, encodé en UTF-8 avec une déclaration XML incluse. **La fonction generer_html_depuis_xml()** convertit le fichier XML contenant les chunks analysés en une page HTML. Elle lit les balises <paragraphe> et <phrase> et affiche chaque chunk à l'aide d'un bloc dont la classe CSS est déterminée par la catégorie grammaticale (N, V, ADV). Chaque chunk est encodé sous forme : [N contenu]. Un lien est inséré vers une feuille de style chunking.css qui définit les couleurs, polices et effets visuels associés à chaque type de chunk. Les mots non reconnus (UNK) sont temporairement stockés dans un tampon pour être regroupés lorsqu'ils se suivent pour alléger l'affichage HTML. À la fin de chaque phrase, un saut de ligne
 est inséré pour conserver la structure du texte d'origine.

La fonction principale est **def main()** car elle permet l'enchaînement complet des étapes. Elle commence par charger le lexique et les règles de segmentation à partir de fichiers externes. Elle lit ensuite le texte, remplace les sauts de ligne par le marqueur BR, puis applique une phase de pré-tokenisation qui prépare le texte. Le texte est segmenté en tokens sur lesquels sont appliquées les règles. Les segments reconnus sont formatés sous la forme [catégorie contenu] tandis que les mots non analysés sont laissés tels quels. Le résultat est affiché dans la console. Une fois l'analyse terminée, la fonction génère un fichier XML contenant tous les chunks structurés selon leur nature grammaticale, la règle appliquée, et la structure du texte (paragraphes, phrases). Ce fichier XML est ensuite converti en une page HTML, grâce à une fonction dédiée. **if __name__ == '__main__': main()** permet d'exécuter le script uniquement s'il est lancé directement et non lorsqu'il est importé comme module dans un autre fichier.

Le fichier *chunker.dtd* définit la structure du fichier XML produit par le moteur de chunking. Il impose une organisation hiérarchique dans laquelle l'élément <**texte**> contient un ou plusieurs éléments <**paragraphe**>. Chaque paragraphe est lui-même composé d'une ou plusieurs balises <**phrase**> et chaque phrase contient un ou plusieurs éléments <**chunk**>.

L'élément <**texte**> doit obligatoirement posséder deux attributs : src pour la source du texte et date qui précise la date à laquelle l'analyse a été effectuée. Les chunks contiennent le texte brut à l'intérieur de la balise et ont deux attributs obligatoires : cat pour la catégorie grammaticale attribuée (N, V, ADV..), et regle qui précise quelle règle a permis de produire ce chunk.

Le fichier *chunking.css* définit l'apparence visuelle des chunks affichés dans la page HTML générée à partir du fichier XML. Chaque chunk est affiché dans une balise à laquelle s'applique la classe .chunk. Cette classe utilise une police sans serif, une petite taille de texte (0.65em) pour compacter l'affichage et des bords arrondis. Chaque chunk est présenté sous forme de bloc visuellement distinct. Avant le contenu textuel du chunk, ::before insère automatiquement la catégorie grammaticale entre crochets ([N]) affichée en marron et ::after ajoute un espace après chaque chunk. L'attribut cat permet d'appliquer des styles spécifiques à chaque catégorie grammaticale grâce à des sélecteurs (.chunk[cat="N"], .chunk[cat="V"]). Chaque type de chunk a sa propre couleur pour faciliter l'identification visuelle de sa fonction grammaticale. Les chunks non reconnus (UNK) sont affichés en rouge.

Comparez les résultats de votre chunker avec ceux attendus (texte chunké à la main) de manière quantitative et qualitative. Justifiez/expliquez les erreurs de votre chunker

Sur les 254 chunks manuels, 223 ont été correctement reconnus. Le taux d'erreur global est d'environ 12 %. Il faut distinguer deux types d'erreurs : celles où le chunk n'est pas reconnu du tout (UNK ou mauvaise catégorie) et celles de segmentation, où le chunk est partiellement correct mais n'englobe pas tous les mots attendus. Globalement, la majorité des chunks sont correctement identifiés. La plupart des erreurs relèvent donc de mauvaises segmentations plutôt que de mauvaises catégorisations. Cela signifie que le système détecte la bonne nature des groupes mais n'en saisit pas toujours les limites exactes. Un facteur d'erreur reste l'ordre des règles dans le fichier : très souvent, un chunk échoue non pas par manque de règle, mais parce qu'une règle trop générale passe avant une autre plus précise. Il aurait été simple de corriger ces cas un par un, mais le défi était justement d'expliquer les erreurs sans intervenir directement sur les règles. Il aurait été possible d'atteindre un taux de reconnaissance de 100 % en adaptant finement le lexique ou en écrivant des règles très spécifiques. Toutefois, ce n'était pas l'objectif ici. L'enjeu est de généraliser cette base de règles pour qu'elle reste réutilisable sur d'autres textes en français ou d'autres langues.

Toutes les ponctuations —16 **PCTF**, 20 **PCTNF**, 7 **GO**, 7 **GF** — ont été correctement identifiées, de même que les 6 **BR** (marqueurs de retour à la ligne). Cela montre que la reconnaissance des séparateurs syntaxiques est maîtrisée. Les 16 conjonctions de coordination (**COORD**), le **PDEM**, ainsi que les 2 **PRELSUJ** ont également été correctement reconnus. Il en va de même pour le **PVANT**. Ce bon résultat s'explique par la présence explicite de ces éléments dans le lexique (par exemple, *GO => [GO* ou *PRelSuj => [PRelSuj*) et par l'efficacité de règles précises comme **ir => [VER* pour les infinitifs ou *P + *ant => [PVANT* pour les formes en -ant. Les **ADV** (5 attendus et 7 trouvés) et les **ADJ** (6 attendus et 10 trouvés) ont été

surgénéré. La règle *ique => Adj a bien fonctionné. Mais elle a souvent produit des chunks isolés (des adjectifs seuls) alors que dans la plupart des cas, un adjectif n'a pas de sens seul et devrait être rattaché à un groupe nominal ou adverbial. Pour les ADV, beaucoup ont été identifié par le lexique et la règle ADV => [ADV, alors que certains se trouvaient dans un chunk verbal ou nominal.

Concernant les groupes prépositionnels nominaux (**PN**) et même si leur reconnaissance soit satisfaisante quantitativement (45 chunks manuels pour 55 détectés), on observe un phénomène de sur-segmentation : certaines règles trop générales (comme P + Det + _) englobent des éléments qui ne devraient pas l'être, allongeant les chunks. Un comportement similaire est observé pour les groupes nominaux (**N**), avec 49 attendus et 40 correctement identifiés. Ici encore, les erreurs tiennent à une mauvaise gestion des frontières syntaxiques, les règles ayant tendance à englober trop d'éléments. Les règles comme Det + _ + _ => N ou Det + _ posent un problème selon leur priorité. Il y a aussi beaucoup d'erreur sur les structures verbales **V** (25 manuels, 19 détectés, 6 erreurs), **SV** (12 manuels, 7 détectés, 5 erreurs) et **Pver** (13 manuels, 10 détectés, 3 erreurs). On peut envisager la création d'une règle *ent => V pour capturer certains verbes au présent. Toutefois, cela risque d'introduire du bruit car de nombreux adverbes se terminent également en *-ent. Aucune règle ne couvre les formes *re => Ver ou P + *re => Pver, ce qui laisse de côté certains infinitifs importants. Les structures négatives (avec *n'... pas* ou juste *n'*) sont mal reconnues. La règle de type PPS + Neg + _ => SV manque souvent de flexibilité pour englober les compléments.

Pour freiner	Pver
la fonte	N
des glaces	PN
,	Pctnf
le gouvernement	N
fait installer	V
3000 climatiseurs	N
sur la banquise	PN
	BR

Voici l'analyse des erreurs phrases par phrases pour plus de précision. *[Pver Pour freiner] [N la fonte] [PN des glaces] [PCTNF ,] [N le gouvernement] [V fait installer] [N 3000 climatiseurs] [PN sur la banquise] [BR]* Dans le titre de l'article, le système identifie correctement les 9 chunks attendus, soit une exactitude de 100 %. *Face [PN à l'urgence] [Adj climatique] [PCTNF ,] [N le gouvernement] [V a décidé] [Pver d'agir] vite [Coord et] [ADV bien] [Pvant en annonçant] [N l'installation] [ADV de plus] [PN de trois mille] climatiseurs [PN sur la banquise] [PCTF .]* En revanche, dans la deuxième ligne, seuls 12 chunks sur les 14 attendus sont reconnus correctement, donnant un taux d'exactitude d'environ 86 %. D'abord, la séquence « Face à » aurait dû être reconnue comme un chunk de type Lp, mais elle ne l'est pas car aucune règle spécifique n'a été définie pour ce cas particulier. Ensuite, « à l'urgence » est capturé en tant que PN, ce qui est incorrect : la préposition « à » est incluse à tort avec le groupe nominal. Le mot « climatique » de son côté, est bien catégorisé comme adjectif grâce à la règle *ique => [Adj mais il est séparé du nom « urgence » auquel il se rattache. Cette segmentation vient de l'application directe de la règle sans tenir compte du contexte nominal environnant. Enfin, dans l'expression « de trois mille climatiseurs », le mot « climatiseurs » n'est pas intégré au chunk prépositionnel (PN). Cela est dû à la règle P + _ + _ => [PN, qui est volontairement limitée à deux éléments après la préposition pour éviter de capturer des groupes trop longs. Cette limitation empêche ici la reconnaissance complète du groupe nominal. Concernant le groupe verbal, le chunk Pver est correctement détecté pour « d'agir ». Mais l'adverbe « vite » qui suit immédiatement, n'est pas

Face à	Lp
l'urgence climatique	N
,	Pctnf
le gouvernement	N
a décidé	V
d'agir vite	Pver
et	Coord
bien	Adv
en annonçant	PVant
l'installation	N
de plus	Adv
de trois mille climatiseurs	PN
sur la banquise	PN
.	Pctf

englobé dans le même chunk. Pourtant, la règle P + *ir + _ => [Pver] existe, suggérant que le système ne prend pas suffisamment en compte de mots postposés dans ce type de structure.

Reportage réalisé [PN avec le WWF] [PCTF .][BR] Pour la 3e ligne, 3 des 5 chunks sont correctement identifiés. Le mot « reportage » n'est pas reconnu comme un nom (N) car il est absent du lexique. Ce choix est volontaire : en l'absence de déterminant, son identification comme nom aurait été trop spécifique. De même, « réalisé » n'est pas catégorisé comme verbe (V) faute de règle *é => [V.

Reportage	N
réalisé	V
avec le WWF	PN
.	Pctf
	BR

[PN Dès l'été] prochain [PCTNF ,] ce [V sont] [ADV plus] [PN de 3000 climatiseurs] [PRelSuj qui] vont [V être largués] [PN sur la banquise] [PN par un immense] pont aérien [PCTF .] La quatrième ligne affiche 8 chunks correctement identifiés sur les 9 attendus. L'expression « dès l'été prochain » n'est pas entièrement reconnue comme un groupe prépositionnel nominal (PN) : le mot « prochain » reste isolé. Cette segmentation résulte de la règle P + Det + _ => [PN qui ne prend en compte que deux éléments après la préposition. De manière similaire dans « par un immense pont aérien », le chunker s'arrête à « par un immense » et ne reconnaît pas « pont aérien » en raison de la même limitation de portée. Par ailleurs, la séquence verbale « vont être largués » n'est pas traitée comme un ensemble : seul « être largués » est identifié comme un verbe par la règle ModInf + _ => [V, qui couvre également deux mots. Enfin, « ce » n'est pas reconnu comme pronom à cause d'un conflit dans le lexique où il aurait été à la fois identifié comme déterminant et pronom, imposant un choix.

Dès l'été prochain	PN
,	Pctnf
ce sont plus	SV
de 3 000 climatiseurs	PN
qui	PRelSuj
vont être largués	V
sur la banquise	PN
par un immense pont aérien	PN
.	Pctf

[GO «] [N L' objectif] [V est] [Pver de refroidir] durablement [N la banquise] [Pver pour éviter] [CSUB qu'] [SV elle ne fonde] davantage [PCTNF ,] voire même qu'elle se reconstitue [PN en partie] [Coord ou] totalité [GF »] [V a expliqué] [N un porte-parole] [PN du gouvernement] [PCTF .] Dans cette cinquième phrase, 19 chunks étaient attendus, et le chunker en reconnaît correctement 16. Le mot « davantage » est resté hors chunk alors qu'il aurait dû faire partie du groupe verbal « elle ne fonde davantage ». La règle PPS + Neg + _ => [SV ne permet pas son inclusion car l'ajout d'un ADV dans cette configuration serait trop spécifique pour être intégré au lexique, et une règle plus large comme PPS + Neg + _ + _ => [SV introduirait trop de bruit. Par ailleurs, même si la règle ADV + CS => [CSUB soit bien définie et que les éléments « voire même » (ADV) et « qu' » (CSUB) soient présents dans le lexique, le chunker ne parvient pas à identifier cette séquence. Cela s'explique probablement par le fait que « voire même » est enregistré comme une seule unité avec un espace dans le lexique, alors qu'il apparaît en deux mots distincts dans le texte. La structure verbale « elle se reconstitue » n'est pas reconnue comme un chunk verbal (SV) malgré la présence des formes « elle » (PPS) et « se » (Pref) dans le lexique. De plus, le segment « [V est] [VInf de refroidir] durablement » est mal segmenté : le chunk attendu était [V est] suivi de [Pver de refroidir durablement], l'adverbe « durablement » devant faire partie du Pver. L'erreur semble venir d'une mauvaise interprétation de la règle P + *ir + _ => [Pver. Enfin, le mot « totalité » n'a pas été intégré dans un chunk nominal alors qu'il était annoté comme [N totalité] dans la référence manuelle. N'étant

L'objectif	N
est	V
de refroidir durablement	Pver
la banquise	N
pour éviter	Pver
qu'	Csub
elle ne fonde davantage	SV
,	Pctnf
voire même qu'	Csub
elle se reconstitue	SV
en partie	PN
ou	Coord
totalité	N
»	GF
a expliqué	V
un porte-parole	N
du gouvernement	PN
.	Pctf

pas présent dans le lexique et apparaissant sans déterminant, le système n'a aucun moyen de le reconnaître comme nom.

[GO «] [PN Avec ce largage] [PN de 1 climatiseur] [N tous les 100] kilomètres carrés [PCTNF ,] [N le dérèglement] [Adj climatique] [V sera] [PN de l'histoire] ancienne [GF »] a-t-il ajouté [PCTNF ,] affirmant qu'il s'agissait là [PN de la seule] solution réalisable [PN d'un point] [PN de vue économique] [Coord mais] [ADV surtout] [Adj écologique] [PCTF .] [BR] A la sixième ligne, 21 chunks étaient attendus, et le système en a correctement identifié 17. Le groupe « tous les 100 kilomètres carrés » est partiellement reconnu : seuls les trois premiers mots sont identifiés comme un groupe nominal [N] tandis que « kilomètres carrés » reste en dehors. Cela semble liée à la règle Det + Det + _ => [N, qui ne prend pas en compte un quatrième élément. De même dans le groupe « de l'histoire ancienne », l'adjectif « ancienne » n'est pas inclus dans le groupe prépositionnel nominal [PN] attendu. Une autre lacune concerne la structure verbale inversée « a-t-il ajouté », qui n'est pas reconnue comme [SVInv] : la difficulté à interpréter « a-t-il » comme auxiliaire inversé empêche l'activation de la règle Mod + *-PPS => [SVInv. Le groupe verbal subordonné « affirmant qu'il s'agissait là » est ignoré : aucun chunk [Vant], [CSUB] ou [SV] n'est formé. L'expression « de la seule solution réalisable » n'est pas entièrement regroupée : « solution réalisable » est exclu du chunk nominal car la règle [P + _ + _ => PN] ne couvre que trois éléments. Les expressions adjectivales « de vue climatique » ou « surtout écologique » sont mal interprétées : elles sont analysées uniquement comme adjectifs (ADJ) alors qu'elles auraient dû faire partie d'un groupe nominal (PN) ou adverbial (ADV). Cela s'explique par la règle *ique => [ADJ] qui prédomine dans l'analyse.

«	GO
Avec ce largage	PN
de 1 climatiseur	PN
tous les 100 kilomètres carrés	N
,	Pctnf
le dérèglement climatique	N
sera	V
de l'histoire ancienne	PN
»	GF
a-t-il ajouté	SVInversé
,	Pctnf
affirmant	Vant
qu'	Csub
il s'agissait là	SV
de la seule solution réalisable	PN
d'un point	PN
de vue économique	PN
mais	Coord
surtout écologique	Adj
.	Pctf
	BR

[N Le gouvernement] [V a cependant] rappelé [CSUB que] [PN d'autres options] restaient [PN sur la table] [PCTF .] 7 chunks étaient attendus selon l'annotation manuelle, et le chunker en a correctement identifié 6. Tout d'abord, dans le groupe verbal « a cependant rappelé », le verbe au participe passé « rappelé » n'est pas inclus dans le chunk. Cela s'explique par l'absence de règle permettant de combiner un modale, un adverbe et un verbe, de type Mod + _ + _ => [V] ou Mod + ADV + _ => [V], même si « cependant » est bien présent dans le lexique. Ensuite, le verbe « restaient » n'est pas reconnu comme un chunk verbal. En l'absence d'un pronom sujet (PPS) le précédant, et sans règle spécifique prenant en compte les désinences en -ent, le système ne dispose d'aucun moyen pour l'identifier comme un verbe isolé.

Le gouvernement	N
a cependant rappelé	V
que	Csub
d'autres options	PN
restaient	V
sur la table	PN
.	Pctf

[GO «] [SV Nous avons mobilisé] [N nos meilleurs] cabinets [PN de conseil] [PN sur le sujet] [PCTNF ,] [N la banquise] [V sera sauvée] [PCTNF ,] [PPS vous] n' [V avez] [ADV plus] rien [PN à craindre] [GF »] [PCTF .] Ici, 13 chunks étaient attendus selon l'annotation manuelle, et le système en identifie correctement 11, soit un taux de reconnaissance de 84,6%. Premièrement, la structure nominale « nos meilleurs cabinets » est mal segmentée : seul le groupe « nos meilleurs » est identifié comme [N] tandis que « cabinets » est laissé isolé. Cette erreur provient de l'absence d'une règle de type Det + Adj + N => [N qui

«	GO
Nous avons mobilisé	SV
nos meilleurs cabinets	N
de conseil	PN
sur le sujet	PN
,	Pctnf
la banquise	N
sera sauvée	V
,	Pctnf
vous n'avez plus rien	SV
à craindre	Pver
.	Pctf
»	GF

permettrait de fusionner correctement les trois éléments en un chunk nominal complet. Deuxièmement, dans l'exemple « vous n'avez plus rien », l'analyse ne parvient pas à regrouper l'ensemble de la séquence en un chunk verbal cohérent de type [SV], contrairement à l'annotation manuelle. Le système segmente: « vous » en [PPS], « n' » non reconnu, « avez » en [V], « plus » en [ADV], et « rien » en UNK. La principale cause est la non prise en compte de la négation combinée à la règle ADV => [ADV qui priorise l'étiquetage isolé de l'adverbe. Enfin, l'expression « à craindre » est à tort analysée comme un groupe nominal [PN] alors qu'un groupe verbal prépositionnel [Pver] était attendu. Même si une règle de type P + *re => [Pver puisse permettre cela, elle risquerait d'introduire trop de bruit en s'appliquant à des cas non pertinents.

[N Le gouvernement] [V a annoncé] aussi [Ver réfléchir] [Pver à doubler] [Coord ou] [Ver tripler] [N le nombre] [PN de climatiseurs] [PCTNF ,] construire [N un très] gros réfrigérateur [PN dans lequel] [SV on placerait] [N la banquise] [N l' été] [CSUB quand] [SV il fait] trop chaud [Coord ou] [Ver installer] [PN des très grosses] pyramides [PN de pains] [PN de glace] [PCTF .] Dans cette phrase, 23 chunks étaient attendus selon l'annotation manuelle, et le système en reconnaît correctement 21. Dans « [PN des très grosses] pyramides », le nom « pyramides » est exclu du chunk. Là encore, la limitation du nombre d'éléments pris en compte dans la règle empêche une construction complète du groupe nominal. Le verbe « construire » n'est pas reconnu comme [V], en l'absence d'une règle du type *re => [V] qui permettrait de capturer les infinitifs en -re. L'adverbe « aussi » identifié dans le lexique comme ADV n'est pas intégré au chunk verbal « [V a annoncé] ». Cela s'explique par le fait que la règle Mod + _ => [V] ne prend pas en compte les ADV postposés. Enfin, la segmentation de « [N un très] [UNK gros réfrigérateur] » échoue en raison d'un mauvais ordre de priorité entre les règles. De même, l'expression « trop chaud » n'est pas reconnue, car l'adverbe « trop » est absent du lexique.

Le gouvernement	N
a annoncé aussi	V
réfléchir	Ver
à doubler	Pver
ou	Coord
tripler	Ver
le nombre	N
de climatiseurs	PN
,	Pctnf
construire	Ver
un très gros réfrigérateur	N
dans lequel	PN
on placerait	SV
la banquise	N
l'été	N
quand	Csub
il fait	SV
trop chaud	Adj
ou	Coord
installer	Ver
des très grosses pyramides	N
de pains de glace	PN
.	Pctf

[GO «] [N La dernière] option [V sera] [PN de faire] [PN de grandes collectes] [PN de glace] [CSUB lorsque] [N les Français] dégivrent [N leur frigo] [Coord et] [PN de les rejeter] [ADV ensuite] [PN sur la banquise] [GF »] [PCTF .] [BR] Ici, 17 chunks étaient attendus selon l'analyse manuelle, et le chunker en reconnaît correctement 14. Une première erreur concerne le groupe nominal « [N La dernière] option » où l'article et l'adjectif « La dernière » sont segmentés ensemble tandis que le nom « option » est laissé à part. Ce découpage incorrect met en évidence une limite dans l'application des règles : la structure Det + _ + _ => [N n'est pas priorisée, et seule la règle plus restreinte Det + _ est activée. Ensuite, les séquences verbales « de faire » et « de les rejeter » ne sont pas correctement analysées. Elles sont interprétées à tort comme des groupes prépositionnels nominaux [PN] en raison de la règle P + _ => [PN faute de règles spécifiques comme P + *re => [Pver] ou P + Pref + *er => [Pver]. La présence du mot « les » qui peut être à la fois déterminant et pronom dans le lexique, compliquerait davantage l'analyse. Enfin, le verbe « dégivrent » reste isolé en tant qu'UNK car

«	GO
La dernière option	N
sera	V
de faire	Pver
de grandes collectes	PN
de glace	PN
lorsque	Csub
les Français	N
dégivrent	V
leur frigo	N
et	Coord
de les rejeter	Pver
ensuite	Adv
sur la banquise	PN
.	Pctf
»	GF
	BR

aucune règle de type *ent => [V] n'est définie pour identifier les formes verbales conjuguées en -ent.

[N Le gouvernement] [V a critiqué] [N les détracteurs] [PN de ce projet] [PRelSuj qui] pointent [PN du doigt] [N un gaspillage] [Adj énergétique] [Coord et] [N une pollution] massive [PCTF .] 11 chunks étaient attendus selon l'annotation manuelle, et 10 ont été correctement reconnus par le système, soit un taux de reconnaissance de 91%. Le verbe « pointent » est resté isolé en tant qu'UNK faute d'une règle du type *ent => [V] permettant de l'identifier comme verbe conjugué. Par ailleurs, dans le groupe nominal « une pollution massive », l'adjectif « massive » n'a pas été intégré au chunk [N]. De plus, l'adjectif « énergétique » est isolé en tant que [ADJ], sans lien avec un groupe nominal.

Le gouvernement	N
a critiqué	V
les détracteurs	N
de ce projet	PN
qui	PRelSuj
pointent	V
du doigt	PN
un gaspillage énergétique	N
et	Coord
une pollution massive	N
.	Pctf

«	GO
Une fois	N
de plus	Adv
,	Pctnf
ces gens	N
sont	V
dans une posture politique	PN
,	Pctnf
nous sommes	SV
les seuls	N
à proposer	Pver
des solutions	N
logiques	Adj
et	Coord
simples	Adj
à mettre	Pver
en place	PN
»	GF
a martelé	V
le porte-parole	N
du gouvernement	PN
.	Pctf
	BR

[GO «] [N Une fois] [ADV de plus] [PCTNF ,] [N ces gens] [V sont] [PN dans une posture] [Adj politique] [PCTNF ,] [PPS nous] [V sommes] [N les seuls] [Pver à proposer] [PN des solutions logiques] [Coord et] simples [PN à mettre] [PN en place] [GF »] [V a martelé] [N le porte-parole] [PN du gouvernement] [PCTF .] [BR] 23 chunks étaient attendus selon l'annotation manuelle, et le système en a correctement identifié 19. Le mot « politique » n'est pas intégré dans le groupe nominal attendu, il est traité comme un adjectif isolé. De même, la séquence « nous sommes » est incorrectement segmentée en deux chunks distincts — [PPS nous] et [V sommes] — alors qu'une règle de type PPS + _ => [SV est bien définie et aurait dû s'appliquer ici. Le mot « logiques », généralement interprété comme adjectif, devait cette fois faire partie du groupe nominal « des solutions logiques », mais il a été analysé de façon incorrecte. L'adjectif « simples » n'est pas reconnu, faute d'entrée dans le lexique et d'aucune règle permettant son intégration. Enfin, l'expression « à mettre » est analysée à tort comme un groupe prépositionnel nominal [PN], alors qu'un groupe verbal prépositionnel [Pver] était attendu.

[BR] Face [PN à ce genre] [PN de solutions farfelues] [PCTNF ,] [N le WWF] tient [Pver à rappeler] [CSUB que] [N la vraie] lutte [PN contre le dérèglement] [Adj climatique] ne se fera pas avec [PN des climatiseurs] [Coord mais] avec [PN des actions concrètes] [PCTF .] 15 chunks étaient attendus selon l'annotation manuelle, et 11 ont été correctement identifiés. « Face » n'est pas pris en compte, car aucune règle ne traite les locutions prépositionnelles (LP), ce qui entraîne une segmentation incorrecte de « [PN à ce genre] » au lieu du groupe nominal attendu « [N ce genre] ». Le verbe « tient » n'est pas reconnu comme tel en l'absence d'une règle spécifique permettant d'identifier cette forme conjuguée. De même, le mot « lutte » est exclu du groupe nominal « la vraie lutte ». Concernant la séquence verbale négative « ne se fera pas avec », le chunker ne distingue pas correctement les éléments du groupe verbal, la négation n'étant pas prise en

Face à	Lp
ce genre	N
de solutions farfelues	PN
,	Pctnf
le WWF	N
tient	V
à rappeler	Pver
que	Csub
la vraie lutte	N
contre le dérèglement climatique	PN
ne se fera pas	V
avec des climatiseurs	PN
mais	Coord
avec des actions concrètes	PN
.	Pctf

compte. Par ailleurs, la préposition « avec » n'est pas intégrée au groupe nominal qui suit (« des climatiseurs »). De plus, « des » est ici étiqueté uniquement comme déterminant (Det) dans le lexique, et non comme préposition (P), ce qui empêche l'application de cette règle — une révision du lexique pourrait donc être envisagée même si cela pose la question de l'ambiguïté si « des » est considéré comme P.

[GO «] [N La préservation] [PN des océans] [PCTNF ,] [N la lutte] [PN contre le braconnage] [Coord et] [N l'accélération] [PN de la transition] [Adj écologique] [V sont autant] [PN de vrais combats] [CSUB que] [SV nous devons] [Ver mener] [GF »] précise [N un porte-parole] [PN de l'organisation][PCTF .] [BR] 19 chunks étaient attendus selon l'annotation manuelle, et 17 ont été correctement identifiés. Le mot « écologique » est analysé isolément au lieu d'être intégré au groupe prépositionnel nominal « de la transition écologique » à cause d'une règle insuffisamment étendue pour inclure les adjectifs postposés dans un [PN]. La séquence verbale « nous devons mener » n'est pas reconnue comme un seul chunk [SV] ; elle est découpée en [V devons] et [Ver mener] traduisant un découpage dû à la priorité accordée à certaines règles sur d'autres. Enfin, le verbe « précise » n'est pas identifié comme tel : il est absent du lexique, et aucune règle ne permet de le classer en [V].

«	GO
La préservation	N
des océans	PN
,	Pctnf
la lutte	N
contre le braconnage	PN
et	Coord
l'accélération	N
de la transition écologique	PN
sont autant	V
de vrais combats	PN
que	Csub
nous devons mener	SV
»	GF
précise	V
un porte-parole	N
de l'organisation	PN
.	Pctf
	BR

[Coord Et] [Pver pour agir] efficacement [PCTNF ,] [Ver soutenir] [PN des programmes cohérents] [Coord et] [PREL où] [N les solutions] [V ne viennent pas] [PN d'un manuel] [Pver d' électroménager] [PCTNF ,] faites [ADV plutôt] [N un don] [PN au WWF] [PCTF .] 16 chunks étaient attendus selon l'annotation manuelle, et 14 ont été correctement identifiés. L'adverbe « efficacement » n'est pas inclus dans le groupe verbal prépositionnel [Pver], en raison d'une absence de règle permettant l'insertion d'un adverbe à cette position. De plus, le mot « d'électroménager » est analysé comme un [Pver] au lieu du [PN] attendu, à cause de la règle P + *er => [Pver]. Le mot « faite » n'est pas identifié comme un verbe, ce qui est cohérent avec l'état actuel du chunker : il est absent du lexique et aucune règle ne permet sa reconnaissance en [V]. Cependant, la structure verbale négative « ne viennent pas » est bien reconnue comme un chunk verbal [SV] ce qui montre que la gestion des négations est parfois correctement traitée.

Et	Coord
pour agir efficacement	Pver
,	Pctnf
soutenir	Ver
des programmes cohérents	N
et	Coord
où	PRrel
les solutions	N
ne viennent pas	V
d'un manuel	PN
d'électroménager	PN
,	Pctnf
faites plutôt	V
un don	N
au WWF	PN
.	Pctf

[GO «] [ADV Alors] oui [PCTNF ,] [CSUB peut-être que] [N notre idée] n' [V est] pas aussi cool [CSUB que] [PDEM celle] [Pver de larguer] 3 000 climatiseurs [PN sur la banquise] [PCTNF ,] [Coord mais] on [SV vous promet] [CSUB que] [PCTNF ,] niveau efficacité [PCTNF ,] [SV c' est tout] [PN de même] [ADV mieux] [GF »] [PCTF .] 23 chunks étaient attendus selon l'annotation manuelle, et 20 ont été correctement identifiés, soit un taux de reconnaissance de 87 %. Dans l'expression « alors oui », le mot « oui » n'est pas pris en compte, ce qui est attendu puisqu'aucune règle n'étend les ADV à ce type d'interjection affirmative. Les éléments de négation « n' » et « pas » ne sont pas correctement reconnus ensemble. L'adjectif modifié « aussi cool » n'est pas identifié comme un groupe adverbial, en

«	GO
Alors oui	Adv
,	Pctnf
peut-être que	Csub
notre idée	N
n'est pas	V
aussi cool	Adj
que	Csub
celle	Pdem
de larguer	Pver
3 000 climatiseurs	N
sur la banquise	PN
,	Pctnf
mais	Coord
on vous promet	SV
que	Csub
,	Pctnf
niveau efficacité	N
,	Pctnf
c'est tout de même	SV
mieux	Adv
.	Pctf
»	GF

partie parce que « cool » n'est pas dans le lexique en tant qu'ADV. De plus, la quantité « 3 000 climatiseurs » n'est pas reconnue correctement. Bien que « 3 000 » figure dans le lexique, la présence de l'espace entre les chiffres entraîne une segmentation défaillante contrairement à l'analyse correcte de « 3000 » en début d'article. La structure « on vous promet » n'est pas analysée en bloc car « on » est ignoré, faute d'une règle de type ProS + Pro + _ => [SV] qui permettrait de regrouper les pronoms sujet et objet avec le verbe. De même, l'expression « niveau efficacité » est mal analysée : les deux noms à la suite ne peuvent être intégrés à un même chunk sans que les deux soient explicitement présents dans le lexique, ce qui serait trop spécifique. Enfin, la structure « c'est tout de même » est mal segmentée. Le mot « même » est classé à tort comme [PN] à cause de la règle P + _ => [PN].

»	GF
Et	Coord
donc	Coord
Le Gorafi	N
soutient	V
le WWF	N
,	Pctnf
et	Coord
vous	PPS
?	Pctf

[Coord Et] [Coord donc] [N Le Gorafi] soutient [N le WWF] [PCTNF ,] [Coord et] [PPS vous] [PCTF ?] 9 chunks étaient attendus selon l'annotation manuelle, et 8 ont été correctement identifiés, soit un taux de reconnaissance de 91%. Le verbe « soutient » n'est pas reconnu comme tel, à cause de l'absence d'une règle de type *ent => [V], ce qui empêche son identification en tant que verbe conjugué.

Sélectionnez un autre texte, de même source, même langue et même typologie que le premier et analysez les résultats. Quelles sont les modifications à apporter au lexique, à la base de règle, voire au moteur pour améliorer les résultats

L'autre texte s'intitule «*Écologie – Pour les protéger des pailles en plastique, Donald Trump va supprimer les tortues* », nommé texte2.txt. Tout comme le texte précédent, les ponctuations — soit **PCTF**, **PCTNF**, **GO**, **GF** — ont été correctement identifiées, de même que **BR** (marqueurs de changement de paragraphe). Cela montre que la reconnaissance des séparateurs syntaxiques est maîtrisée. Les conjonctions de coordination (**COORD**), le **PDEM**, ainsi que les 2 **PRELSUJ** ont également été correctement reconnus. Les groupes prépositionnels nominaux **PN** bien que largement détectés (33 identifiés), comportent 5 erreurs souvent dues à des règles trop restrictives (limitées à deux ou trois mots) pour « *des mers du globe*, *à l'aide de huit stylos* ». Les groupes nominaux **N** ont un taux d'erreur plus important avec 15 correctement détectés sur 20, à cause de l'absence de noms propres dans le lexique comme *un Donald Trump*. Les verbes **V** montrent également des faiblesses : 14 identifiés pour seulement 10 corrects. Les participes passés (*stockées*, *traquées*) ou les subjonctifs (*soit imprimé*) ne sont pas reconnus, faute de règles comme *ées => [V ou Mod + Part => [V. Les chunks **[SV]** ne sont correctement détectés que dans 2 cas sur 4, souvent à cause de l'absence de pronoms sujets comme *j'*, *elles*, ou de la non-reconnaissance des modaux. Aucun chunk **[SVInv]** n'a été formé, alors qu'au moins une inversion « *a-t-il poursuivi* » était attendue.

Du côté des groupes verbaux infinitifs **Pver**, la moitié seulement (3 sur 6) ont été correctement identifiés, à cause de limites dans les règles comme P + *er => [Pver], qui

échouent si des adverbes ou pronoms s'intercalent. Les infinitifs **Ver** sont mieux traités (3 corrects sur 3) grâce à une règle efficace *er => [Ver]. Les groupes en participe présent **Pvant** posent aussi problème : sur 3 détectés un seul est correct, car les adverbes postposés ne sont pas englobés (*en les délocalisant intégralement*). Enfin, les catégories **ADJ** et **ADV** sont globalement bien reconnues (2 à 3 erreurs) mais elles souffrent parfois d'un lexique insuffisant (*exempt*, *recyclé*, *intégralement*, *fébrile*) ou de l'absence de règles permettant leur rattachement à des noms ou verbes.

On constate que très souvent, dès qu'un mot est absent du lexique, il est soit ignoré soit mal catégorisé. Les règles seront aussi soit trop courtes à certains endroits soient trop longues. Les balises peuvent aussi être parfois bloquantes s'il elles se trouvent entre les groupes nominaux ou prépositionnels. Ce moteur de règles n'est pas assez robuste, il ne prend pas en compte la variation lexicale et syntaxique. Son usage va rester limité aux textes de même style, de même typographie et même source (et encore !).

Écologie – [PN Pour les protéger] [PN des pailles] [PN en plastique] [PCTNF ,] Donald Trump [V va supprimer] [N les tortues] [BR] 7 chunks sont correctement identifiés sur les 9 attendus. Le mot « Écologie » n'est pas reconnu comme un nom en raison de l'absence de déterminant devant lui et de son absence du lexique. De plus, « Donald Trump » n'est pas interprété comme un nom propre : aucune règle ni entrée lexicale ne permet de regrouper les deux mots en une unité car la catégorie [NP] (nom propre) n'avait pas été nécessaire dans le traitement du premier texte et n'existe donc pas encore dans le chunker. Il faudrait ajouter une règle Maj + Maj => [NP ou les intégrer séparément dans le lexique.

Après [Ver avoir] annoncé [N le retour] [PN des pailles] [PN en plastique] [PCTNF ,] [N le président] Donald Trump [V a souhaité] faire un [GO «] petit geste [PN pour la planète] [GF »] [Pvant en supprimant] purement [Coord et] simplement [N les tortues] [PN de l'ensemble] [PN des mers] [PN du globe] [PCTF .] Reportage [PCTF .] 14 chunks sont correctement identifiés sur les 21 attendus. Le mot *« après »* n'est pas reconnu comme une préposition : il reste isolé en [UNK] car il est absent du lexique. De même, l'expression *« avoir annoncé »* est incorrectement fragmentée. Seul *« avoir »* est identifié comme un [Ver], tandis que *« annoncé »* participe passé, n'est ni chunké ni relié à un groupe verbal. Cela résulte de l'absence de règle du type ModInf + *é => [V. Le groupe *« des pailles en plastique »* est lui aussi découpé de manière inadéquate : les éléments sont séparés en deux chunks [PN], alors qu'un unique chunk [PN des pailles en plastique] serait attendu. Cela s'explique par une règle PN limitée à deux mots après la préposition, ce qui empêche la capture d'un groupe plus long. Un allongement des règles PN serait envisageable. Un autre problème concerne la structure verbale *« a souhaité faire »* qui devrait être traitée comme un bloc verbal unique. Actuellement, seul *« a souhaité »* est reconnue comme [V] tandis que *« faire un »* est rejeté en [UNK. L'ajout de règles permettrait de mieux gérer ces cas. Le groupe nominal *« un petit geste »* est également mal analysé. L'introduction de la citation directe avec [GO «] interrompt le chunk, ce qui empêche le système de reconnaître l'unité syntaxique. Une meilleure gestion des balises de citation permettraient d'éviter ce type de rupture. Les adverbes *« purement »* et *« simplement »* sont ignorés car absents du lexique. Ils sont donc logiquement rejetés en [UNK]. Il serait essentiel de les ajouter avec la

catégorie ADV. Enfin, le mot *« Reportage »*, bien qu'attendu comme [N], est laissé hors chunk. Cela découle à la fois de l'absence de déterminant et de son absence du lexique.

[BR] C' [V est] [N un Donald] Trump visiblement ému [PRelSuj qui] s' [V est adressé] [N ce mercredi] matin [PN à la presse] depuis [N le bureau] Ovale [PN de la Maison-Blanche] [PCTF .] [GO «] J' ai [V vu hier] [Ver soir] [N un reportage] [PN sur National Geographic] montrant [N l' impact] [PN des pailles] [PN en plastique] [PN sur la faune] marine [PCTF .]
Sur les 23 attendus, 15 sont correctement détectés. Tout d'abord, « C' est » est incorrectement séparé : *« C' »* n'est pas reconnu comme pronom sujet, empêchant la formation du chunk [SV]. De même, « Donald Trump » est scindé : *« Trump »* n'est pas intégré au nom [N un Donald], faute de règle ou de lexique pour les noms propres. L'expression « visiblement ému » est ignorée alors qu'elle correspond à une séquence adjectivale attendue comme [ADJ]. La forme pronominale « s'est adressé » est mal analysée : *« s' »* n'est pas pris en compte ce qui empêche le regroupement verbal. Le groupe « ce mercredi matin » est mal traité : *« matin »* est isolé alors qu'il devrait être intégré dans le [N]. Ensuite, le mot « depuis » absent du lexique, reste hors chunk. Le chunk « [N le bureau] Ovale » est incomplet : l'adjectif *Ovale* est laissé de côté alors qu'il devrait être englobé. Plus loin, « J'ai vu » n'est pas reconnu comme [SV] car *« j' »* n'est pas inclus dans les PPS. L'erreur sur « soir » classé en [Ver], provient d'une surapplication de la règle *ir => [Ver] qui devrait être restreinte. Le participe présent « montrant » reste hors chunk, faute de règle [Vant] pour les formes verbales non précédées d'une préposition. De plus, « des pailles en plastique » est segmenté en deux [PN] au lieu d'un seul, alors qu'un regroupement plus long serait approprié. Enfin, dans « sur la faune marine », l'adjectif *marine* est exclu du chunk [PN]. Il faudrait ajouter « j' » au lexique respectivement comme PPS et il est aussi nécessaire d'intégrer « Donald » et « Trump » comme noms propres ou de créer une règle Maj + Maj => [NP] pour les regrouper automatiquement. La règle *ir => Ver doit être restreinte car elle provoque des erreurs comme pour « soir ». Il faut également enrichir le lexique en adjectifs et adverbes et étendre les règles PN et N pour capturer des groupes comme *des pailles en plastique* ou *le bureau ovale*. Enfin, une règle *ant => [Vant] permettra de reconnaître correctement *montrant*.

[SV C' est terrifiant] ! [GF »] a-t-il expliqué [PCTNF ,] fébrile [PCTNF ,] [Pvant en aspirant] trois litres [PN de sirop] [PN de maïs] [PN à l' aide] [PN d' une paille] [PN en plastique] [PCTF .] [GO «] Voilà pourquoi j' ai décidé [PN de faire] [PN des océans] [N un endroit] [ADV plus] sûr [PN pour nos tortues] [PN en les délocalisant] intégralement [PN dans le Sahara] [PCTNF ,] sans possibilité [PN de retour] [GF »] [PCTNF ,] a-t-il poursuivi [Pvant en signant] frénétiquement [N le décret] [PN à l' aide] [PN de huit stylos] Bic [PCTF .]
Le point d'exclamation n'est pas identifié, bien qu'il marque une ponctuation forte. L'inversion « a-t-il expliqué » n'est pas reconnue comme [SVInv], faute d'activation de la règle dédiée (Mod + *-PPS => [SVInv]). Le mot « fébrile » n'est pas catégorisé comme adjectif alors qu'il devrait être intégré au lexique. Le groupe « trois litres de sirop de maïs » est fragmenté et « d'une paille en plastique » devrait former un chunk unique, ce qui nécessite une règle [PN] plus large. Les mots « voilà » et « pourquoi » ne sont pas reconnus comme [ADV] : ils devraient être ajoutés au lexique. La structure « j'ai décidé » n'est pas analysée comme [SV] car « j' » n'est pas dans les PPS. De plus, « de faire » est traité comme un [PN] alors qu'un [Pver] est

attendu. L'expression « plus sûr » est séparée et « en les délocalisant intégralement » est mal chunkée : seul *en les délocalisant* est reconnu, alors que l'adverbe *intégralement* devrait y être intégré. Le groupe « sans possibilité » attendu en [PN] est ignoré, et à nouveau « a-t-il poursuivi » n'est pas traité comme [SVInv]. Enfin, « frénétiquement » est exclu du chunk verbal alors qu'il devrait être absorbé dans [Pvant], et « Bic » est rejeté car au-delà de la limite de deux éléments après une préposition dans les règles PN. Pour améliorer ces phrases, il faut d'abord ajouter au lexique les formes manquantes. Côté règles, il est essentiel de corriger Mod + *-PPS => [SVInv] pour les inversions comme *a-t-il expliqué*. Il faut aussi étendre les règles PN à quatre éléments pour capturer *d'une paille en plastique* ou *de huit stylos Bic*. Pour les verbes, on ajoutera des règles du type *ant + ADV => [Pvant] afin d'englober les adverbes dans les chunks verbaux.

[BR] [N L'opération] [PCTNF ,] [PN d'envergure] [PCTNF ,] [V devrait débuter] [PN d'ici] [N les prochaines] semaines [Coord et] [Ver mobiliser] plusieurs dizaines [PN de milliers] [PN de bateaux] [PN de pêche] [Coord et] [PN de torpilleurs japonais] [PN sur l'ensemble] [PN du planisphère] [PCTF .] [N Les tortues] [PCTNF ,] traquées "mortes [Coord ou] vives" [PCTNF ,] seront [PN par la suite] stockées [PN en fond] [PN de cale avant] d' [V être catapultées] [PN par millions] [PN en plein milieu] [PN du Sahara] depuis [N la soute] [PN des avions] [PN de chasse] [PN de l'US] [Ver Air] Force [PCTF .] Le groupe « les prochaines semaines » est partiellement reconnu : *« semaines »* est isolé, faute de règle Det + Adj + N => [N]. De même, « plusieurs dizaines de milliers » et « de bateaux de pêche » devraient être regroupés chacun en un seul chunk [PN] (ce que les règles actuelles à deux ou trois éléments ne permettent pas). Le mot « traquées » participe passé, reste hors chunk car il manque une règle de type *ées => [V]. Les adjectifs « mortes » et « vives » devraient être catégorisés en [ADJ] et ajoutés au lexique. Le verbe « seront » est absent du lexique en tant que modal (MOD), ce qui empêche sa reconnaissance en [V]. Le participe « stockées » n'est pas chunké faute d'une règle adaptée (Mod + *ées => [V]). La structure « avant d' » devrait être traitée comme un tout, mais elle est fragmentée. Le mot « depuis » n'est pas dans le lexique comme préposition, ce qui empêche la construction du [PN] suivant. Enfin, les groupes « des avions de chasse » et « de l'US Air Force » devraient être regroupés mais sont mal analysés : le système traite *Air* comme un verbe en *ir, à tort (faute d'une exception dans la règle *ir => [Ver]). Pour améliorer, il faut ajouter au lexique les mots *seront* (Mod), *mortes*, *vives* (ADJ), *depuis* (P), ainsi que US *Air Force* comme NP.

[GO «] [N Le Sahara] [V est] [N un lieu] exempt [PN de tout plastique] [PRelSuj qui] [N leur garantira] [N un havre] [PN de paix] [PREL où] elles pourront enfin vivre heureuses [Coord et] apprendre à se reconstruire [GF »] [PCTNF ,] [V a commenté] [N le président] [PN de la première] puissance mondiale [PCTNF ,] avant [Pver de réaffirmer] sa volonté de [GO «] prendre possession [PN des océans] [Pvant en remplaçant] [N l'ensemble] [PN des mers] [PN du globe] [PN par une gigantesque] station balnéaire [PN avec vue] [PN sur la mer] [GF »] [PCTF .] Le mot « exempt » attendu en [ADJ], reste isolé car il est absent du lexique. Le pronom « elles », manquant lui aussi, empêche la formation du chunk [SV] *elles pourront enfin*. De plus, « vivre » n'est pas reconnu comme [Ver] à cause de l'absence d'une règle *re => [Ver], et « vivre heureuses » est mal analysé. L'infinitif « apprendre » est isolé en [Ver] mais

l'expression « à se reconstruire » qui devrait être [Pver], n'est pas reconnue faute de règle P + Pref + *re => [Pver]. Le groupe « de la première puissance mondiale » est fragmenté alors qu'il devrait être un chunk [PN] complet. Le mot « avant » utilisé comme préposition, est absent du lexique. Enfin « de prendre possession », qui devrait être un [Pver], est mal analysé à cause de l'interruption par [GO «], et les groupes « des mers du globe », « station balnéaire », « avec vue », et « sur la mer » sont segmentés à tort, faute de règles PN suffisamment longues.

[BR] [N Une posture] [GO «] green-friendly [GF »] [PN pour le président] [Coord et] [GO «] écocidaire [GF »] [PN pour les associations] [PCTNF ,] [PRelSuj qui] semble n' [V être qu'un] début [PCTNF ,] puisque Donald Trump s' [V est également] engagé à ce [CSUB que] chacun [PN des décrets anti-écologiques] qu'il signera [PN à l' avenir] soit imprimé sur [PN du papier] 100 % recyclé [PCTF .] Les mots « green-friendly » et « écocidaire » ne sont pas reconnus car ce sont des néologismes ou emprunts non présents dans le lexique. La structure verbale « semble n'être » est mal analysée. Le mot « puisque » utilisé comme conjonction est absent du lexique et devrait être ajouté comme CS. L'expression « à ce que » devrait être capturée comme un tout, mais il n'existe pas de règle. « qu'il signera » est ignoré, et « soit imprimé » n'est pas chunké correctement. Enfin, « 100 % recyclé » échappe à l'analyse, car le système ne gère pas les chaînes contenant un pourcentage.

Sélectionnez un autre texte, d'une source et d'une typologie différente, mais dans la même langue et réitérez l'expérience et l'analyse subséquente

Le texte3.txt est une des *Lettres écrites de la montagne* de Jean-Jacques Rousseau ABU - TEXTE lettresecrites1. Sur environ 160 chunks attendus, seuls 117 ont été détectés automatiquement, dont 94 sont jugés corrects. Cela signifie qu'environ 66 chunks ont été mal segmentés ou complètement ignorées.

Exemple : *[BR] C'est [Ver revenir] tard [PCTNF ,] je [N le sens] [PCTNF ,] [PN sur un sujet] trop rebattu [Coord et] déjà presque oublié [PCTF .] Mon état [PCTNF ,] [PRelSuj qui] ne me permet [ADV plus] aucun travail suivi [PCTNF ,] mon aversion [PN pour le genre] [Adj polémique] [PCTNF ,] ont causé ma lenteur [PN à écrire] [Coord et] ma répugnance [Pver à publier] [PCTF .] J'aurais même tout [PN à fait supprimé] [N ces Lettres] [PCTNF ,] [Coord ou] [ADV plutôt] je lie [N les aurais] point écrites [PCTNF ,] s'il n'eût été question [CSUB que] [PN de moi :] [Coord Mais] ma patrie [V ne m'est pas] tellement devenue étrangère [CSUB que] je puisse [Ver voir] [V tranquillement opprimer] ses citoyens [PCTNF ,] [ADV surtout] lorsqu'ils n'ont compromis leurs droits qu'en défendant ma cause [PCTF .] Je serais [N le dernier] [PN des hommes si] [PN dans une telle] occasion j'écoutais [N un sentiment] [PRelSuj qui] n'est [ADV plus] ni douceur ni patience [PCTNF ,] [Coord mais] faiblesse [Coord et] lâcheté [PCTNF ,] [PN dans celui qu'il] empêche [Pver de remplir] son [Ver devoir] [PCTF .]*

Le moteur à base de règle détecte toujours aussi bien les ponctuations Pctf, Pctnf, GO, GF et les retours à la ligne BR sur ce texte littéraire, tout comme les conjonctions de coordination COOR et de subordination CSUB, ainsi que les pronom relatif sujet PRelSuj. Les

groupes nominaux N et prépositionnels nominaux PN sont globalement bien détectés sauf s'ils sont trop longs ou complexes. Plusieurs groupes nominaux prépositionnels sont découpés en plusieurs [PN] alors qu'ils devraient être reconnus comme un seul ensemble. Cela s'explique par des règles trop restrictives, limitées à deux ou trois éléments. *« de tout honnête homme »* est scindé, il en va de même pour *« dans les rues de Syracuse »* divisé à tort en deux [PN]. En parallèle, la forte densité d'adjectifs et d'adverbes complexifie l'analyse. Beaucoup restent isolés alors qu'ils devraient être intégrés à des chunks existants : *« un langage glacé »* ou *« travail suivi »*. D'autres expressions comme *« si fort irrités »* ou *« tellement devenue étrangère »* combinent adverbe et participe, mais ne sont pas reconnues à cause de l'absence de règles Pour améliorer les performances, il faudrait allonger les règles [PN] à 4 ou 5 éléments, intégrer les adjectifs postposés dans les [N], et enrichir le lexique avec des adjectifs/adverbes plus littéraires.

Une des erreurs fréquentes observées concerne la reconnaissance des verbes conjugués [V] et des structures verbales [SV] notamment dans des formulations littéraires comme *« s'il n'eût été question »*, *« ont si fort irrités trouveront »*, *« eût enflammé mon cœur »* ou encore *« Archimède tout transporté courait nu »*. Le moteur de règles actuel est conçu pour dans un français contemporain standard, et peine à traiter les formes verbales moins courantes ou plus soutenues. Pour l'améliorer, il serait nécessaire d'enrichir le lexique avec des formes conjuguées du subjonctif, du passé simple ou du conditionnel, et de développer des règles plus flexibles. Le système ne reconnaît pas certaines ponctuations comme *« ! »* ou *« : »* car elles sont absentes du lexique.

Plus largement, le bon fonctionnement de ce chunker repose sur un lexique très complet et une grande précision dans la formulation des règles. Pour garantir de bonnes performances sur tout type de texte, il faudrait préalablement enrichir le lexique avec un maximum de formes, y compris les variantes grammaticales et orthographiques. Cela pose la question des doublons et des ambiguïtés, car de nombreux mots peuvent remplir plusieurs fonctions grammaticales selon le contexte.

Sélectionnez un texte simple dans une autre langue, adaptez le lexique et réitérez l'expérience et l'analyse subséquente

Le texte4.txt est « El tiempo », tiré du site Lingua.com. C'est un court récit narratif et descriptif en espagnol, destiné à un public apprenant la langue. Il s'agit d'un texte de niveau débutant à intermédiaire écrit dans un registre simple, conçu pour familiariser avec le vocabulaire lié à la météo et aux phénomènes naturels. El tiempo - Texte espagnol

Le lexique extrait du texte « *El tiempo* » couvre principalement des catégories grammaticales fondamentales pour structurer un discours simple en espagnol. Il inclut d'abord une série de déterminants (Det) comme *el, la, los, las, un, una*. On trouve aussi des ponctuations (*., !, ,*) classées comme *Pctf* ou *Pctnf*. Il y a les conjonctions de coordination (CC) telles que *y, pero, o* et les conjonctions de subordination (CS) (*que, aunque, si, pues, porque*). Les prépositions (P) (*a, de, en, con, por, para*) sont très fréquentes. Les pronoms personnels sujets

(PPS) comme *me* et *nos* ainsi que *se* en *ProS* (forme pronominale ou réfléchie sont présent, ainsi que la négation (no). Plusieurs formes conjuguées de verbes auxiliaires (*ha, han, habrá, es, está, están, estuvo, será*) sont identifiées comme *Mod*. Des adverbes courants (ADV) comme *hoy, allí, aquí* marquent le temps et le lieu.

Exemple : *[N El tiempo] [BR] [ADV Hoy] hace [N mucho frío] [PCTF .] [V Es invierno] [Coord y] [N todas las calles] [V están cubiertas] [PN de nieve] [PCTF .] Dentro [PN de poco vendrá] [N la primavera] [Coord y] [PN con ella] [N el sol] [Coord y] [N el tiempo] cálido [PCTF .] [N La semana] pasada [V estuvo] [PN de lluvia] [Coord y] tormenta [PCTF .] Incluso [N un rayo] cayó encima [PN de la campana] [PN de la catedral] [PCTNF ,] [Coord pero] no ocurrió nada [PCTF .] [N Los truenos] siempre [SV me han dado] miedo [Coord y] [N mucho respeto] [PCTF .] [Coord Pero] tenemos suerte [PCTF .] [PCTF .] [PCTF .] [CSUB pues] [N la previsión] del tiempo [PN para mañana] [V es muy] buena [PCTF .] Dicen [CSUB que] [ADV hoy] [V habrá heladas] [Coord y] [PN por la tarde] granizo [PCTNF ,] [Coord pero] mañana [N el día] [V será soleado] [PCTF .] [Pver A ver] [CSUB si] tengo suerte [Coord y] veo [N algún arcoíris] [PCTF .]*

J'ai repris les mêmes catégories grammaticales que pour le lexique français afin d'analyser le texte espagnol. De manière générale, le chunker a correctement segmenté les unités, avec une répartition globalement satisfaisante. Toutefois, plusieurs erreurs persistent. Du côté des verbes, certaines formes conjuguées ne sont pas reconnues à cause de l'absence de règles adaptées. Un autre problème propre à l'espagnol est l'utilisation de signes de ponctuation inversés (comme ¿ ou ¡). Ils sont souvent collés au mot suivant et non pris en compte car absents du lexique et mal prétraités. Certains adverbes sont également ignorés ou rejetés comme cela avait déjà été observé dans le traitement du texte en français. Enfin, les irrégularités de conjugaison en espagnol entraînent encore des échecs de reconnaissance.

L'espagnol, tout comme le français, est une langue flexionnelle : les terminaisons verbales y indiquent la personne, le temps, le mode et parfois le nombre. Cela explique pourquoi une fois le lexique adapté, le système de règles utilisé pour le français continue de fonctionner relativement bien en espagnol même sans modification. Les deux langues partagent une syntaxe de type SVO (sujet-verbe-objet). Une différence importante réside dans le fait que l'espagnol est une langue dite *prodrop* : il est fréquent d'omettre le pronom sujet. Celui-ci est implicite et identifiable grâce à la terminaison du verbe. Dans *hablo* (je parle), le pronom *yo* n'est pas nécessaire.

De surcroit, le moteur à base de règles conçu pour le français ne fonctionnerait pas efficacement sur l'allemand par exemple. L'ordre des mots très variable : le verbe conjugué occupe la deuxième position en phrase principale mais est rejeté à la fin en subordonnée. De plus, les déclinaisons sont omniprésentes avec quatre cas (nominatif, accusatif, datif, génitif) qui influencent la forme des articles, rendant inefficace un simple repérage basé sur des déterminants fixes. Enfin, la formation courante de mots composés très longs (comme *Krankenhausverwaltung*) échapperait au chunker.

CONCLUSION

Le moteur à base de règles fonctionne de manière satisfaisante sur l'article du Gorafi, mais montre ses limites sur d'autres genres comme les écrits littéraires. L'analyse d'un extrait des *Lettres écrites de la montagne* de Rousseau révèle que de nombreux chunks échappent à la détection. La segmentation y est moins régulière, le lexique plus dense et les constructions plus complexes.

L'intégration d'un texte en espagnol, langue flexionnelle à syntaxe similaire (SVO), confirme également les limites du chunker. Malgré un bon taux de reconnaissance générale, certaines formes verbales conjuguées ne sont pas reconnues. Les signes de ponctuation inversés propres à l'espagnol ou encore la possibilité d'omettre le sujet pronominal (langue pro-drop) ne sont pas non plus pris en charge.

Si un moteur fondé sur des règles peut fonctionner localement sur un corpus spécifique, il reste inadapté dès qu'on le confronte à des types de textes, des styles ou des langues différents. Même en enrichissant le lexique et en allongeant les règles, la couverture ne sera jamais totale. Cela confirme qu'un moteur fondé sur des règles fixes ne peut prétendre couvrir la diversité du langage naturel.