

Marina Seghier P3 – 21705105
Eirini Rozalia Ompasogkie P10 – 41006700
Doruntina Fazliu P10 – 38011906
Regina Costa P10 – 41000639

Projet de recherche : Problèmes de mathématique du 3ème cycle

L'objectif de ce projet était de détecter des indices d'opérations, des expressions temporelles et des entités nommées difficiles à comprendre ou ambiguës dans des problèmes de mathématiques de 3ème cycle. Pour les indices d'opérations, nous avons écrit un code python, et pour les expressions temporelles et les entités nommées, nous avons utilisé un logiciel de graphes : Unitex.

→ Le corpus

Le corpus d'origine comptait 1409 énoncés, mais présentait de nombreux doublons. Comme nous voulions procéder à une annotation manuelle avant de passer aux annotations automatiques (Unitex et python), nous avons supprimé les doublons grâce à une fonctionnalité dédiée sur Google Sheets. Suite à cette suppression, nous sommes passées de 1409 à 1334 énoncés que nous nous sommes répartis en 4 (soit 334 ou 333 chacune) afin d'annoter les différentes expressions temporelles, les entités nommées et les opérations mathématiques de nos énoncés.

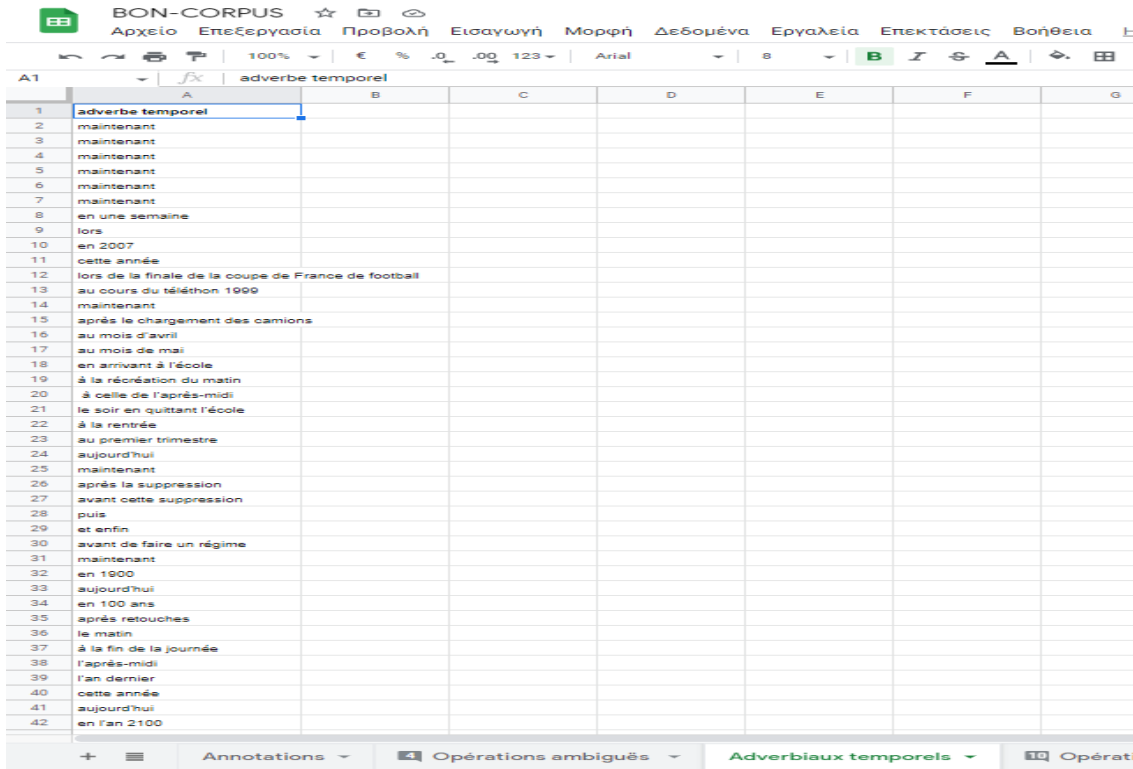
BON-CORPUS										
Αρχείο Επεξεργασία Προβολή Εισαγωγή Μορφή Δεδομένα Εργασία Επεκτάσεις Βοήθεια										
H τελευταία τροποποίηση πραγματοποιήθηκε στις 3/26/21, από τον χρήστη Doruntina Fazliu										
100% 123										
A1										
	A	B	C	D	E	F	G	H	I	J
305			303 A l'occasion de son mariage, les amis de Vanessa lui offrent un service de table composé de 24 assiettes plates CM			Vanessa	offrent, composé, valeur			
306			304 Un négociant en vin attend sa livraison : - 12 colis contenant chacun 24 bouteilles de bordeaux ; - 8 colis contiennent CM				livraison, contenant, chacun, déjà retenues, mettre en vente			
307			305 Dominique achète sa voiture à crédit. Il verse 1 915 € à la commande, 3 859 € à la livraison de son véhicule et 5 CM			Dominique	achète, verse, commande, livraison, solde, 30 mensualités, revient			
308			306 Un touriste de fond prépare son plan d'entraînement annuel. Lundi : 10 km ; mardi : 20 km ; mercredi : 10 km ; je CM				distance, parcourt, repos, durant, quatre semaines			
309			307 La population mondiale était d'environ 2 500 000 personnes 10 000 ans avant Jésus-Christ. Il y a 2 000 ans, elle CM			avant Jésus-Christ, il y a 2000 à Jésus-Christ	était, multipliée par, 220 fois plus nombreux, multipliée			
310			308 Une salle de cinéma comporte 10 rangées de 12 fauteuils. Les séances ont lieu à 19h, 19h 30, 19h 50. Ce mercredi CM			ce mercredi, dans la journée	compte, rangées, fauteuils, remplie, vendues, ont vu			
311			309 Une famille de cinq personnes part aux sports d'hiver pour une semaine. Elle dépense par personne et par jour 4 CM				dépense, par personne, par jour, pour la semaine, revient			
312			310 Monsieur Walter achète les tomates qu'il a vendues pour payer sa maison : - 25 000 € au début des travaux ; - 4 CM			durant neuf années, chaque mo	tomates, vendues, chaque mois, revenue			
313			311 Le propriétaire d'un terrain vend celui-ci en le divisant en deux lots non constructibles d'une surface respecti CM				vend, en le divisant, surface respective, vente de l'ensemble, rapporte			
314			312 Une conserverie expédie 400 petites boîtes de foie gras de canard à 7 € l'une et 800 boîtes de confit de co CM				l'une, prix de vente, l'ensemble			
315			313 Un représentant de commerce parcourt environ 150 km par jour et travaille en moyenne 20 jours par mois. l'idée CM				parcourt, par jour, 20 jours par mois, un mois de vacances, par an, calculer			
316			314 En un mois, un gars peut cueillir et déposer plus loin environ 4 500 glands. Calcule le nombre de glands ainsi tra CM			chaque jour	cueillir, déposés, transportés, en moyenne, chaque jour pour un mois de 30 jours			
317			315 Madame Némour achète un service de ventes composé de 12 ventes à 10 €, 12 ventes à 20 € et 12 ventes à 30 € CM				service, composé, soldes, prix moyen			
318			316 Un moulinet bat des ailes 12 000 fois par minute. Combien d'ailes représente-t-il de battements par seconde ? CM				par minute, par seconde			
319			317 Le nettoyage d'un animal soigné par le vétérinaire d'une clinique revient environ à 25 €. Combien d' CM				revient, si elle a recueilli			
320			318 Un fût contient 500 L. Combien de bouteilles de 75 cl peut-on remplir avec le contenu de ce fût ? CM				contient, remplir avec le contenu			
321			319 Une piscine olympique mesure 50 m de long. Laure s'entraîne chaque jour en parcourant 5 000 m. Combien de CM			chaque jour	de long, en parcourant, longueurs			
322			320 Une bouteille d'huile contient 180 mL. Combien est-ce de litres de bouteilles ? CM				contient, douzaine			
323			321 Sur une cantonnette, on doit charger des colis pesant chacun 42 kg. La cantonnette ne peut pas transporter plu CM				charger, pesant, chacun, plus de, pour, emporter			
324			322 Un pigeon ramier a migré de Suède jusqu'en France en 25 jours, accomplissant ainsi une distance de 1 700 km CM			chaque jour	migré, distance, moyenne			
325			323 Un lot comprenant un drap de bain et six serviettes de toilette est vendu 68 €. Le drap de bain valant 26 €, CM				lot, comprenant, vendu, valant, prix			
326			324 Quatre frères et sœur se partagent équitablement un héritage d'une valeur de 160 164 €. Quelle est la part de o CM				partagent, équitablement, part, chacune			
327			325 Avec leur camping-car, la famille Jaro a parcouru 5 733 km durant leurs trois semaines de vacances. Combien d' CM			durant leurs trois semaines de	parcourt, kilomètres, en moyenne			
328			326 Pour acheter un réfrigérateur d'une valeur de 649 €, il est possible de ne payer que 150 € à la commande et le si CM				valeur de NUM, la solde, chaque mensualité			
329			327 On décharge un camion contenant 419 colis avec un engin de manutention qui peut transporter 18 colis à la fo CM			au dernier voyage	combien, de colis, combien, voyage			
330			328 Madame Soumer a acheté trois convecteurs électriques identiques et une perceuse. Elle a payé 269 € pour l'en CM				NUM pour l'ensemble, valant NUM, prix d'un		REVOIR d'un	
331			329 D'une étienne contenant 1 500 L d'eau, un jardinier a déjà retiré 24 seaux d'une capacité de 18 L chacun. Combien CM				Combien, de NUM L, chacun			
332			330 Trois personnes vont au restaurant. Elles prennent toutes le même menu, commandant une bouteille de vin et d CM				prix menu			
333			331 La population, en France métropolitaine, est d'environ 62 200 000 habitants, pour un territoire de 550 989 km². C CM				Quelle, à l'unité près par exposé, la densité*			
334			332 Un ouvrier perçoit un salaire de 1 319 € pour un mois de 140 heures de travail. Quel est son salaire horaire ? CM				salaire horaire			
335			333 La région Aquitaine compte environ 3 170 000 habitants. Elle est composée de cinq départements. Calcule le poj CM				population moyenne, un département			
336			334 Dans une lettre publicitaire, l'abonnement à un journal mensuel est proposé, au choix : pour 6 mois à 15 €, pour CM				chaque cas, prix			
337			335 Au Brésil vivent encore 350 000 indiens appartenant à 215 ethnies* différentes. En moyenne, combien d'indivi CM			encore	individue, une ethnies			
338			336 Une course cycliste se déroule sur un parcours comprenant un tronçon de 24 km, un autre de 36 km et, à l'arrivi CM				longueur, km, un tour de circuit			
339			337 Pour les fêtes de fin d'année, l'union des commerçants de Saint-Marcelin organise une tombola dotée de 16 ap CM			Saint-Marcelin	valeur, appareil photo, un platid polaire			
340			338 Mexico est la 3-ville du monde avec environ 23 200 000 habitants. Marseille est la 298-ville du monde avec env CM			Mexico, Marseille	fois plus peuplée			
341			339 Dans un ascenseur, un panneau indique : « Charge maximale autorisée : 1 200 kg. » Combien de personnes o CM				65 kg, prendre place, ascenseur			
342			340 La Corse est divisée en deux départements. La Corse-du-Sud compte 138 600 habitants pour 4 014 km², la Haut CM				au dixième près par défaut, plus forte densité*, population			
343			341 Une sonde spatiale* a voyagé 5 ans et a parcouru 315 millions de kilomètres pour atteindre un astéroïde. Combien CM			5 ans, chaque année	kilomètres, parcourt			
344			342 Dans les pays développés, un habitant d'une grande ville consomme environ 22 barils de pétrole par an, c'est-à- CM				capacité, baril de pétrole			
345			343 Quand on laisse couler l'eau en se lavant les dents, on peut perdre jusqu'à 20 L d'eau. Quelle quantité d'eau pei CM			en une année, deux fois par jour	eau, gaspiller, une personne			
346			344 Vers 1875, l'expédition du challenger effectuait un périple de trois ans et demi sur les océans et parcourait 127 600 CM			Vers 1875, trois ans, chaque mois,	distance, bateau			
Annotations Opérations ambiguës Adverbiaux temporels Opérations (liste)										

Une fois l'annotation manuelle du corpus faite entièrement, nous avons traité chaque catégorie différemment.

→ Unitex : les graphes

1. Les expressions temporelles

Nous avons annoté toutes les expressions qui se trouvaient dans les phrases. Afin de mieux travailler, nous les avons séparé en en mettant une par ligne.



The screenshot shows the BON-CORPUS software interface. The main window displays a list of temporal expressions in a spreadsheet format. The first column (A) contains the expressions, and the subsequent columns (B-G) are empty. The expressions are listed in rows 1 through 42. The first row is labeled 'adverbe temporel' in the header. The expressions include: maintenant, en une semaine, lors, en 2007, cette année, lors de la finale de la coupe de France de football, au cours du téléthon 1999, maintenant, après le chargement des camions, au mois d'avril, au mois de mai, en arrivant à l'école, à la récréation du matin, à celle de l'après-midi, le soir en quittant l'école, à la rentrée, au premier trimestre, aujourd'hui, maintenant, après la suppression, avant cette suppression, puis, et enfin, avant de faire un régime, maintenant, en 1900, aujourd'hui, en 100 ans, après retouches, le matin, à la fin de la journée, l'après-midi, l'an dernier, cette année, aujourd'hui, en l'an 2100.

A	B	C	D	E	F	G
adverbe temporel						
maintenant						
maintenant						
maintenant						
maintenant						
maintenant						
maintenant						
en une semaine						
lors						
en 2007						
cette année						
lors de la finale de la coupe de France de football						
au cours du téléthon 1999						
maintenant						
après le chargement des camions						
au mois d'avril						
au mois de mai						
en arrivant à l'école						
à la récréation du matin						
à celle de l'après-midi						
le soir en quittant l'école						
à la rentrée						
au premier trimestre						
aujourd'hui						
maintenant						
après la suppression						
avant cette suppression						
puis						
et enfin						
avant de faire un régime						
maintenant						
en 1900						
aujourd'hui						
en 100 ans						
après retouches						
le matin						
à la fin de la journée						
l'après-midi						
l'an dernier						
cette année						
aujourd'hui						
en l'an 2100						

Puis, nous avons sauvegardé la liste au format txt et nous avons enlevé les doublons de la liste avec l'aide de la commande Bash `sort nom_du_fichier.txt | uniq -u`. Comme ça nous pouvions avoir le vrai nombre d'expressions temporelles unique de notre corpus. Cette étape-là consistait juste à voir le nombre total d'expressions temporelles uniques, mais la comparaison des résultats (entre l'annotation manuelle et l'annotation automatique) a été faite sur la liste qui contient les doublons.

```
ireneobasogie@DESKTOP-PRA59GI: /mnt/c/Users/irene/Uni/S2/enrichissement_de_corpus
ireneobasogie@DESKTOP-PRA59GI:~$ cd /mnt/c/Users/irene/Uni/S2/enrichissement_de_corpus/
ireneobasogie@DESKTOP-PRA59GI:/mnt/c/Users/irene/Uni/S2/enrichissement_de_corpus$ wc -l liste_adverbes_temporels.txt
576 liste_adverbes_temporels.txt
ireneobasogie@DESKTOP-PRA59GI:/mnt/c/Users/irene/Uni/S2/enrichissement_de_corpus$
```

Étant donné qu'une ligne correspond à une expression temporelle, nous avons utilisé la commande `wc -l nom_fichier.txt` pour trouver le nombre total d'expressions temporelles (avec doublons).

```
liste_adv - Notepad
File Edit View

maintenant
maintenant
maintenant
maintenant
maintenant
maintenant
en une semaine
lors
en 2007
cette année
lors de la finale de la coupe de France de football
au cours du téléthon 1999
maintenant
après le chargement des camions
au mois d'avril
au mois de mai
en arrivant à l'école
à la récréation du matin
à celle de l'après-midi
le soir en quittant l'école
à la rentrée
au premier trimestre
aujourd'hui
maintenant
après la suppression
avant cette suppression
puis
et enfin
avant de faire un régime
maintenant
en 1900
aujourd'hui
en 100 ans
après retouches
le matin
à la fin de la journée
l'après-midi
l'an dernier
cette année
```

```
ireneobasogie@DESKTOP-PRA59GI: /mnt/c/Users/irene/Uni/S2/enrichissement_de_corpus
ireneobasogie@DESKTOP-PRA59GI: /mnt/c/Users/irene/Uni/S2$ cd enrichissement_de_corpus/
ireneobasogie@DESKTOP-PRA59GI: /mnt/c/Users/irene/Uni/S2/enrichissement_de_corpus$ sort liste_adv.txt | uniq -u
21 jours
5 ans
A présent
Un mois plus tard
cent vingt-cinq mille ans
cette année
en arrivant à l'école
en une année
hier
il y a 18 ans
janvier 2004 et janvier 2014
l'année prochaine
l'été dernier
le 31 juillet
l'année écoulée
à celle de l'après-midi
13 ans après
17 août 2015
1814
1886
1905
2 ans
20 ans
2000
2004
2008
2010
2015
2050
21 décembre à 23h15
24 heures
270 minutes
3 h par jour
3 minutes
3 semaines
30 ans
31,5 ans
36 minutes
37 minutes
5 300 ans
5 minutes
51 ans
52 minutes
5h
6 ans
6 ans de plus
7 heures par jour
```

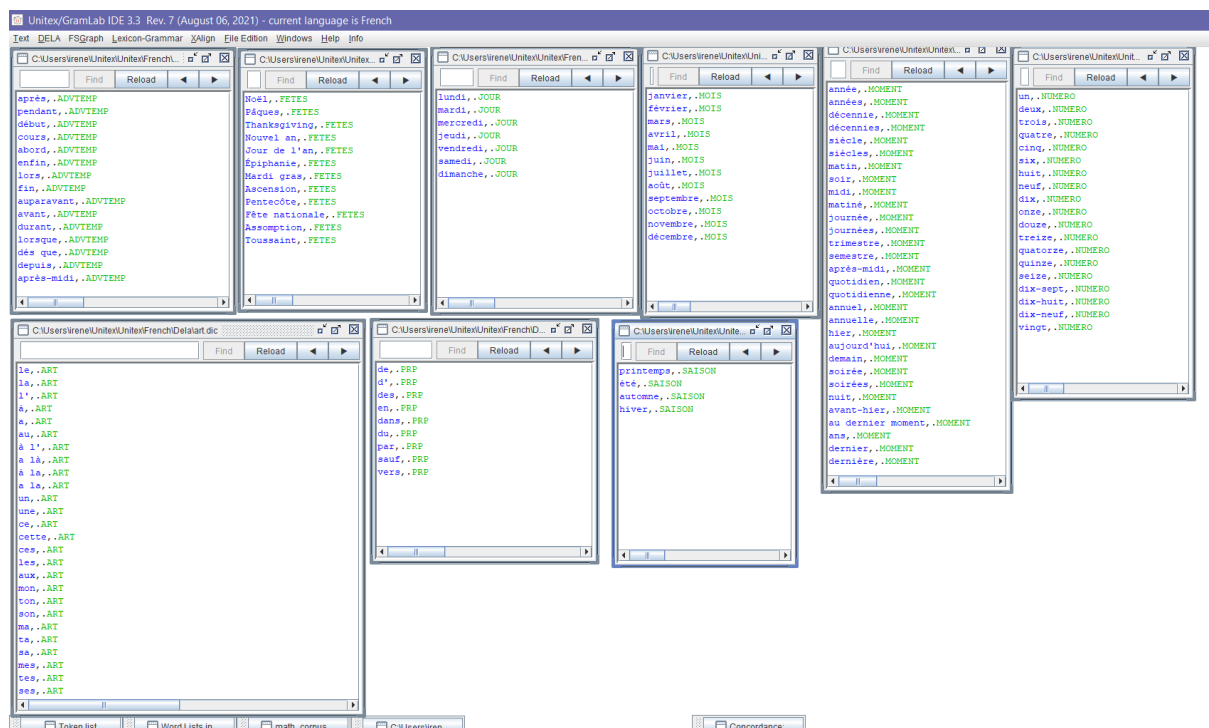
La liste contenait 576 expressions temporelles (avec la suppression de doublons 332).

Pour l'annotation automatique, nous avons dans un premier temps inséré notre corpus dans Unitex au format txt (car il était en format csv avant).

Nous avons choisi le logiciel Unitex et ses graphes pour les raisons suivantes : tout d'abord il est important de souligner qu'Unitex permet la recherche d'expressions complexes et la construction de concordanciers. C'est pour cela qu'il est largement utilisé dans la construction des ressources linguistiques.

Ensuite, tous les objets traités par Unitex sont ou peuvent être transformés en des transducteurs à nombre fini d'états. Dans notre projet, nous nous sommes servis de cette fonctionnalité d'Unitex. Nous avons créé plusieurs graphes qui nous ont permis de reconnaître les entités nommées ambiguës.

Ainsi, nous avons commencé à créer des dictionnaires avec des mots clés. Après la création de dictionnaires, nous voulions créer des graphes pour reconnaître les dictionnaires et nos expressions temporelles.



Les dictionnaires créés.


```

Unitex/GramLab IDE 3.3 Rev. 7 (August 06, 2021) - current language is French
Text DELA FSGraph Lexicon-Grammar XAlign File Edition Windows Help Info
Concordance: C:\Users\irene\Unitex\Unitex\French\Corpus\math_corpus_snf\concord.html
545 matches
ars à 18h 10 et arrive à destination le 2 mars à 1h 15 du matin.{S} Calcule la durée du trajet.
{S} Quel est le prix du blouson ? {S}Le 21 décembre à 23h15, une grand-mère invite ses petits-e
i était programmé sur France 3 le jeudi 3 novembre à 20h35, a débuté avec quatre minutes de ret
43 km au 1 * * juillet, et 13 432 km le 31 juillet.{S} Calcule la distance parcourue par Trista
Adour (Bayonne, Anglet et Biarritz).{S} A 11 h 15, ce jour-là, Marianne décide d'aller pêcher à
rrakech, au Maroc.{S} Son avion décolle à 11 h 45 et arrive à l'aéroport de Marrakech à 13 h 45
idée de la ville.{S} La visite a débuté à 13 h 45 et dure deux heures et dix-sept minutes.{S} A
45 et arrive à l'aéroport de Marrakech à 13 h 45 (heure locale).{S} Quelle heure est-il à Mars
et quatorze minutes.{S} Elle est partie à 13 h 50.{S} A quelle heure finira-t-elle sa promenade
grand-père milanais.{S} Elle a commencé à 14 h 27 et cuisiné pendant trois heures et quarante-h
u.{S} Elle part de Lausanne (en Suisse) à 17 h 30 et arrive à Thonon-les-Bains (en France) à 18
ment lors du troisième arrêt.{S} Enfin, à 17h15, il dépose le reste de ses colis chez son derni
t arrive à Thonon-les-Bains (en France) à 18 h 20.{S} Combien de temps a pris la traversée ? {S
utomobiliste part en voyage le 1er mars à 18h 10 et arrive à destination le 2 mars à 1h 15 du m
ait du sport ? {S}Andréa s'est endormie à 21 h 50.{S} Elle a dormi pendant 10 heures et 16 minu
n part de Paris à 20h et arrive à Dijon à 22h 30.{S} Quelle est la distance entre Paris et Dijo
rd.{S} Sachant qu'il devait se terminer à 22h10, calcule la durée de ce film. {S}Pour aménager
éclipse totale de Soleil s'est achevée à 8 h 01.{S} Elle a duré 3 minutes.{S} A quelle heure l
Léa prend le train à Ambérieu-en Bugey à 8 h 38.{S} La durée de son trajet est de 23 minutes.{
er, madame Mollet a quitté son domicile à 8h 45 pour effectuer une longue randonnée à la vitess
ue enfant reçoit moins de 5 biscuits.{S} A la fin du gouter, il reste 2 biscuits.{S} Combien de
9 contre Cécile.{S} Combien en a-t-elle à la fin ? {S}Ce matin 35 élèves sont absents dans l'éc
ur faudrait-il lire pour finir le livre à la fin des vacances ? {S}En se rendant à la piscine a
entre 2000 et 2009 ?{S} De combien ? {S}A la fin de l'année, la maitresse n'a plus que 29 cahie
en entre l'an dernier et cette année {S}A la fin de l'année, la maitresse n'a plus que 29 cahie
salle de cinéma contient 250 places.{S} A la fin de l'année, cette salle a accueilli 100 000 sp
de kilomètres parcourus par chacun. {S}A la fin de l'été, Tony et sa famille déménagent.{S} So
mbien de kilomètres auront-ils parcourus à la fin de la course ? {S}Benjamin va au cinéma à la s
lesquels elle ajoute 4 kg de sucre.{S} À la fin de la cuisson, qui a duré 1h15, le mélange a p
1 700 tours de manège le matin et 4 326 à la fin de la journée.{S} Combien ont-ils vendu de tou
1 700 tours de manège le matin et 4 326 à la fin de la journée.{S} Combien ont-ils vendu de tou
mes et 15 sacs de 12,5 kg de pommes.{S} A la fin de la journée, 615 kg de pommes ont été cueill
mes et 15 sacs de 12,5 kg de pommes.{S} A la fin de la journée, 615 kg de pommes ont été cueill
este-t-il de billets de 10 € et de 20 € à la fin de la journée dans ce distributeur ? {S}En 187
se rendent dans les grands magasins.{S} À la fin de leurs achats, ils ont dépensé 251 €.{S} Mat

```

Afin de mesurer l'efficacité de nos graphes, nous avons calculé la F-mesure, la précision et le rappel.

	Nb de bons résultats trouvés (automatique)	Nb de résultats à trouver (manuel)		Nb de bons résultats trouvés (automatique)	Nb de résultats trouvés (automatique)
Précision	519	576	Rappel	519	545
	0,9010416667			0,952293578	
F-mesure	0,9259589652				

Le nombre d'expressions temporelles trouvées manuellement est de 576.

Le nombre d'expressions temporelles trouvées automatiquement avec Unitex est de 545.

Et le nombre de bons résultats parmi le total des expressions temporelles trouvées automatiquement avec Unitex est de 519. Pour le dernier, nous avons compté et soustrait les faux résultats. (545 (total de l'annotation automatique) - 26 (faux résultats de l'annotation automatique) = 519 expressions temporelles bien reconnues.

Un exemple de faux résultats est le fait qu'Unitex ait reconnu à partir de notre dictionnaire MOIS.dic la planète "Mars" comme étant le mois "mars".

s à 18h 10 et arrive à destination le 2 mars à 1h 15 du matin. {S} Calcule la durée du trajet. {S} nombre obtiens-tu ? {S} D'octobre 2012 à mars 2013, en France, 43890118 personnes ont joué à un le plus proche du soleil que la planète mars ? {S} Au début du mois, j'avais 1 509 sur mon compte {S} De combien de kilomètres la planète Mars est-elle plus éloignée du soleil que la Terre ? (o e entre les deux champs ? {S} La planète Mars est à 228 millions (ou 228 000 000) de kilomètres itres sont de l'eau salée {S} La planète Mars est éloignée du soleil de 227,9 millions de kilomè soleil se trouve Vénus ? {S} Sachant que Mars est située à 228 000 000 km du soleil, calcule la 328 lettres en février, 200 lettres en mars, et 294 lettres en avril. {S} Combien de lettres a-) Un automobiliste part en voyage le 1er mars à 18h 10 et arrive à destination le 2 mars à 1h 15 de kilomètres. {S} La distance séparant Mars du Soleil varie de 206,7 millions de kilomètres à km du soleil, calcule la distance Terre-Mars. {S} Deux citernes contiennent 500 hl de fioul chac

Un autre problème concernait la reconnaissance des apostrophes. Nous voyons dans les deux captures d'écrans ci-dessous que le "d'abord" avec l'apostrophe correct a été reconnu avec le chemin <PRP> <ADVTEMPS> dans la première ligne.

es ? {S} Résous ce problème en calculant d'abord le prix d'1 kg de cerises. {S} En 4 mois, les ch de temps a-t-elle couru durant le mois d'octobre ? {S} Châtaigne, le chat de Josua, passe en mo ntaines. {S} Quel nombre obtiens-tu ? {S} D'octobre 2012 à mars 2013, en France, 43890118 personn a-t-elle en tout ? {S} Pour le spectacle de fin d'année, la directrice a installé dans le préau nt pas inscrits sur les réseaux sociaux début 2016. {S} Kévin a 830 g de farine et 720 g de sucru'il v avait sur la planète en 2010. {S} Début 2016, on compta 3,025 milliards d'internautes.

Ce qui n'est pas le cas ici :

ne capacité de 2 500 L. {S} Il en tire d'abord 1/5 puis les 3/5 de ce qui reste. {S} Quelle quant litres était pleine. {S} On en soutire d'abord 200 litres puis un quart de ce qui restait. {S} Qu un rouleau de 30 m. {S} Ils utilisent d'abord 3,75 m pour l'entrée de l'école, puis ils confect lection de timbres. {S} Elle en achète d'abord 31. {S} Elle en donne à une amie. {S} Finalement il {S} Samuel verse, dans une bouteille, d'abord deux bols d'une capacité unitaire* de 25 cL, puis altitude inconnue. {S} Le pilote élève d'abord l'avion de 350 m, puis il redescend de 975 m. {S} ? {S} Résous ce problème en calculant d'abord le prix d'1 kg de cerises. {S} En 4 mois, les chev

Nous observons le même problème partout où cette apostrophe apparaît.

oyage d'un animal souillé par le mazout lors d'une marée noire revient environ à 35 €. {S} Combi Calcule le prix total de la voiture. {S} Lors d'un jeu télévisé, un candidat a gagné un lecteur culer à la main le nombre d'équipes. {S} Lors d'une journée exceptionnelle dans une station de s runter si cet achat coûte 16 636 € ? {S} Lors d'un concert dans un stade, on a enregistré 44 435 avant son augmentation de salaire ? {S} Lors d'une compétition de saut en hauteur, un concurren {S} Combien d'argent lui reste-t-il ? {S} Lors d'une course, les cyclistes ont 100 km à parcourir 365. {S} Quel est le nombre d'Edgar ? {S} Lors d'une course automobile, une voiture a parcouru 20 ur) parcourt-il en quatre semaines ? {S} Lors d'un déménagement, on charge, dans une camionnette ourue par l'ensemble des cyclistes ? {S} Lors d'une course cycliste, Kriss effectue 14 tours d'u ervir en achetant les deux briques ? {S} Lors d'une étape du tour de France 2009, le peloton, co d'oiseaux a-t-il compté s en tout ? {S} Lors d'une balade en foret, Essi a ramassé 115 frambo utilise s de litres d'eau en tout ? {S} Lors d'une étape du Tour de France, le peloton, compos s utilisés de litres d'eau en tout ? {S} Lors d'une étape du Tour de France, le peloton, composé s utilisés de litres d'eau en tout ? {S} Lors d'une étape du Tour de France, le peloton, composé 0 % de coton. {S} Quel est le prix ? {S} Lors d'un Grand Prix de formule 1, les pilotes ont 52 t ance les touristes auront-ils parcourue lors de ces deux jours ? {S} Au restaurant, une famille t la distance en km parcourue par Kriss lors de cette compe tition ? {S} J'ai 28 images. {S} Mari m. {S} Quelle distance ont-ils parcourue lors de cette première semaine ? {S} Une annonce publici usines. {S} Combien ai-je de soeurs ? {S} Lors de la finale de la coupe de France de football, on usines. {S} Combien ai-je de soeurs ? {S} Lors de la finale de la coupe de France de football, on plus que lui. {S} Quel âge avait Jacques lors de la naissance de son petit-fils ? {S} La fosse de ien le serveur a-t-il de pourboire ? {S} Lors de leur départ en vacances, les Dupont ont install e taxi note qu'il a parcouru 3 873,2 km lors de ses 23 derniers jours de travail. {S} Trouve la La comète de Halley, découverte en 1682 lors de son passage à proximité de la Terre, revient to lis dans sa camionnette, en dépose sept lors de son premier, repart pour 12 km, puis en dépose

De plus, nous avons rencontré un problème avec le “en” + <NB> qui reconnaît les dates. En dehors de ces dates, le graphe a reconnu beaucoup d’expressions de cette forme-là qui n’étaient pas des expressions temporelles. Nous avons cherché dans le manuel Unix un moyen d’ajouter une contrainte qui nous permettrait de définir que le nombre de <NB> (soit 4 chiffres), mais sans succès.

Cette production est montée à 240 films [en 2008](#). {S} Calcule l’augmentation de la production entre 1900 et 2008 {S} [En 2008](#), entre mai et septembre, on a comptabilisé 576 de la production entre 1900 et 2008 {S} [En 2008](#), en France, il s’est vendu quatre fois plus de extinction. {S} Que peux-tu calculer ? {S} [En 2008](#), en France, 834 000 enfants sont nés, 76 000 mi e atteignait la vitesse de 480 km h. {S} [En 2009](#), son record est de 574,8 km /h. {S} Avec ses 410 il emporte de tuiles sur mon toit ? {S} [En 2009](#), il y avait en moyenne 210 330 voyageurs par jo -il emporté de tuiles sur mon toit ? {S} [En 2009](#), il y avait en moyenne 210 330 voyageurs par jo 'habitants qu'il y avait sur la planète [en 2010](#). {S} Début 2016, on comptait 3,025 milliards d'i , c'est-à-dire 0,42 milliard de plus qu'[en 2010](#). (Source : ONU) Calculer le nombre d'habitants 'oeuvre d'art Chiffres en Vrac réalisée [en 2012](#) par le sculpteur français Fernando Costa. {S} Qu le nombre d'habitants de chaque canton [en 2012](#). {S} Donner un ordre de grandeur du nombre d'hab e 1,95. {S} En 2000, il était de 1,89 et [en 2014](#), il était de 2,01. {S} En quelle année ce nombre rviennent dans cet énoncé ? {S} A Paris, [en 2015](#), un ticket de métro acheté à l'unité coute 1,80 arcourues par Naomi et par Mickaël ? {S} [En 2015](#), la population mondiale était de 7,35 milliards le nombre de personnes vivant en France [en 2016](#) ? {S} En juillet 2015, 3000 personnes ont assist sont rendus dans un parc d'attractions [en 2016](#). {S} En 2017, le parc d'attractions a accueilli dans un parc d'attractions en 2016. {S} [En 2017](#), le parc d'attractions a accueilli 11,9 million aute du monde. {S} Elle mesure 828 m. {S} [En 2019](#), elle sera détrônée par la Kingdom Tower, en co a-t-on de monuments classés, en France, [en 2020](#) ? {S} Dimanche dernier, madame Mollet a quitté s mbien de Français auront plus de 60 ans [en 2050](#). {S} Dans un appartement de 75 m², la salle de b anie. {S} Quel est l'âge de Nicolas ? {S} [En 2050](#), il y aura environ 64 000 000 de Français et un e même robinet ? {S} Un film est projeté [en 24](#) images par seconde. {S} Combien d'images sont proj ramier a migré de suède jusqu'en France [en 25](#) jours, accomplissant ainsi une distance de 1 700 . {S} Combien de bouteilles produit-elle [en 25](#) jours ? {S} Thomas et Florian, qui pèsent respecti rofesseur a découpé un coupon* de tissu [en 28](#) morceaux de 47 cm. {S} Quelle longueur de tissu, e plorateurs arabes, parcourut 120 000 km [en 28 ans](#) de voyages. {S} En moyenne, combien de kilomèt asse à la souris a attrapé 29000 souris [en 29 ans](#). {S} Combien de souris a-t-il attrapées en moy nion forestier a transporté 102 grumes* [en 3](#) voyages. {S} La masse d'un éléphant est environ de sportIf a couru le marathon, 42,195 km, [en 3](#) heures. {S} En moyenne, combien de kilomètres a-t-i s la cheminée. {S} Pour cela, il le scie [en 3](#) morceaux. {S} Quelle sera la longueur des bûches ob une bande de papier de 10,2 cm de long [en 3](#) parties égales. {S} Quelle est la longueur de chaqu ? {S} Buz la sauterelle parcourt 120 cm [en 3](#) sauts. {S} Combien de cm parcourt-elle en un seul s

2. Les entités nommées

Après avoir discuté, 3 grandes classes sémantiques d'entités nommées ont été sélectionnées pour répondre aux besoins des applications du projet MATH. Nous nous sommes concentrées sur les prénoms, les villes (et les régions) ainsi que les pays.

Tout d'abord en ce qui concerne les prénoms, nous avons recueilli dans deux fichiers au format txt des listes de prénoms masculins et féminins les plus communs en français triés par ordre alphabétique. Ces listes ont été trouvées en ligne sur <https://prenom.org/>, en prenant soin de sélectionner l'onglet adéquat (car il y a la possibilité de choisir parmi des entités (prénoms) français, anglais, arabes, espagnols, portugais, turcs et juifs entre autres). Nous avons nettoyé les listes pour obtenir deux fichiers qui contiennent exclusivement les prénoms français les plus communs.

La liste de départ contenait 323 prénoms féminins et 277 prénoms masculins, mais nous avons décidé d'enlever des prénoms désuets, en nous assurant qu'ils n'étaient pas présents dans nos problèmes de mathématiques.

Enfin, la liste des prénoms féminins contenait 308 éléments tandis que la liste de prénoms masculins contenait 257 éléments.

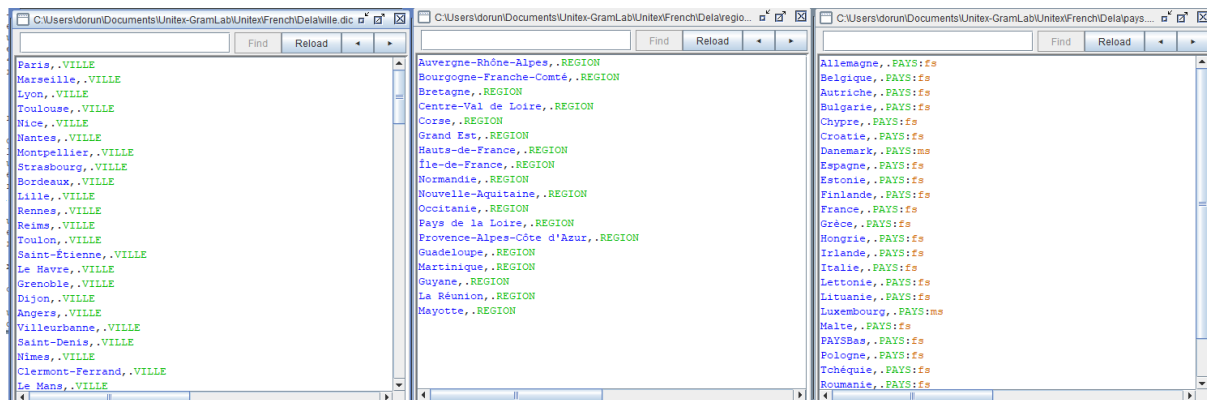
Voici un aperçu de notre liste :

580	Thierry, .N+Hum+PRENOM:ms
581	Thomas, .N+Hum+PRENOM:ms
582	Timothée, .N+Hum+PRENOM:ms
583	Toussaint, .N+Hum+PRENOM:ms
584	Tristan, .N+Hum+PRENOM:ms
585	Ulrich, .N+Hum+PRENOM:ms
586	Urbain, .N+Hum+PRENOM:ms
587	Valentin, .N+Hum+PRENOM:ms
588	Valère, .N+Hum+PRENOM:ms
589	Valéry, .N+Hum+PRENOM:ms
590	Vespasien, .N+Hum+PRENOM:ms
591	Victor, .N+Hum+PRENOM:ms
592	Vincent, .N+Hum+PRENOM:ms
593	Vivien, .N+Hum+PRENOM:ms
594	Xavier, .N+Hum+PRENOM:ms
595	Yanick, .N+Hum+PRENOM:ms
596	Yann, .N+Hum+PRENOM:ms
597	Yannic, .N+Hum+PRENOM:ms
598	Yannick, .N+Hum+PRENOM:ms

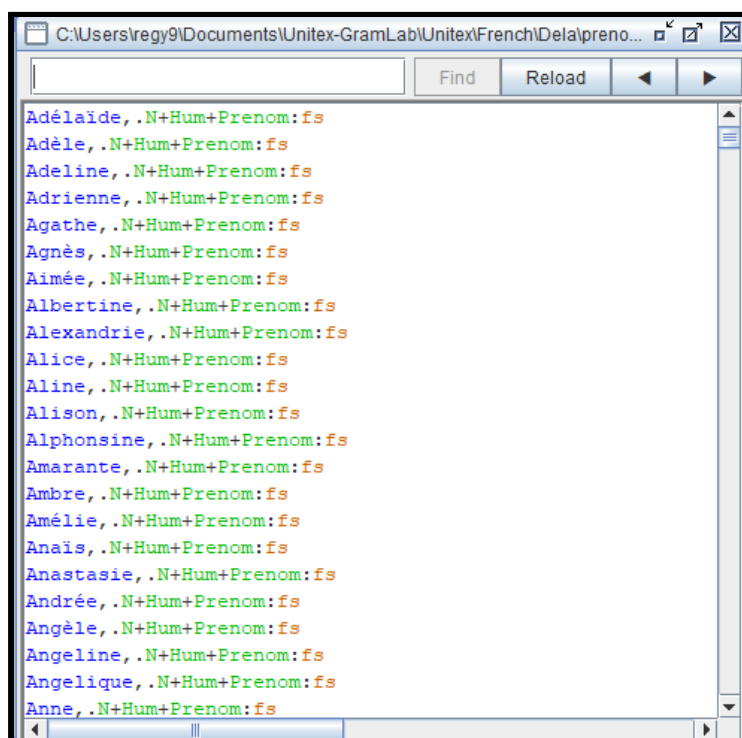
Ensuite, nous avons réalisé que c'était plus pratique de recueillir toutes les entrées dans un même fichier car pour constituer un dictionnaire N+Hum+PRENOM, une seule liste était suffisante. Les marqueurs ms ou fs nous indiquait si le prénom était masculin ou féminin.

Concernant les villes françaises, nous avons pris une liste qui contient les 150 grandes villes françaises. Étant donné que parfois il s'agissait de villes bretonnes, ou de villes plus petites qu'en général nous n'en connaissons pas, et que le but de notre projet est de trouver ce qui peut être ambigu pour des enfants du 3ème cycle, nous avons décidé de nettoyer la liste. Après le nettoyage manuel, nous avons finalement obtenu une liste de 128 villes.

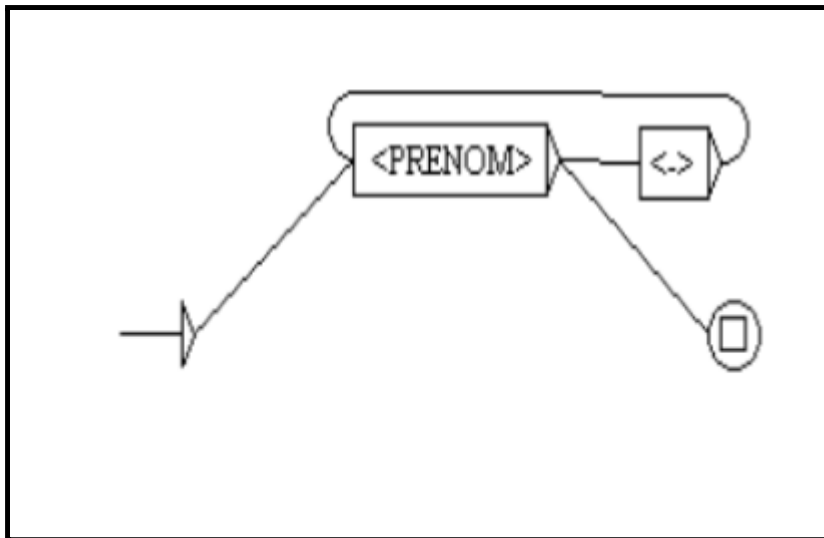
En ce qui concerne la liste des pays, nous avons décidé de prendre les pays de l'Union européenne. Et après avoir remarqué qu'il y avait des régions mentionnées dans le corpus, nous avons pris toutes les régions françaises, y compris les régions d'outre-mer, donc 18 régions au total.



Premièrement, chacune d'entre nous a repéré les entités nommées des catégories correspondantes : prénoms, pays, villes et régions. Puis, nous avons créé et compressé des dictionnaires.



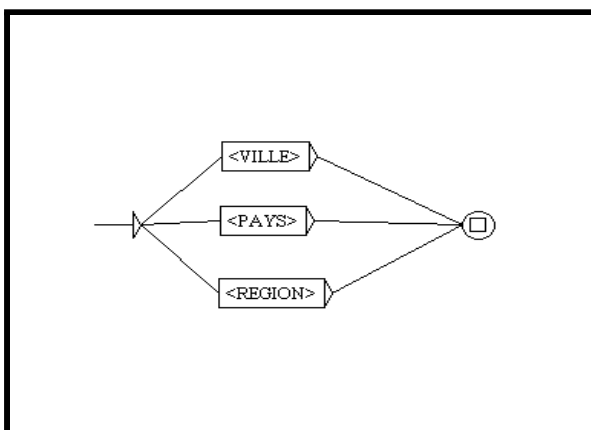
Ensuite, nous avons créé un automate fini qui vise à reconnaître les patrons souhaités.



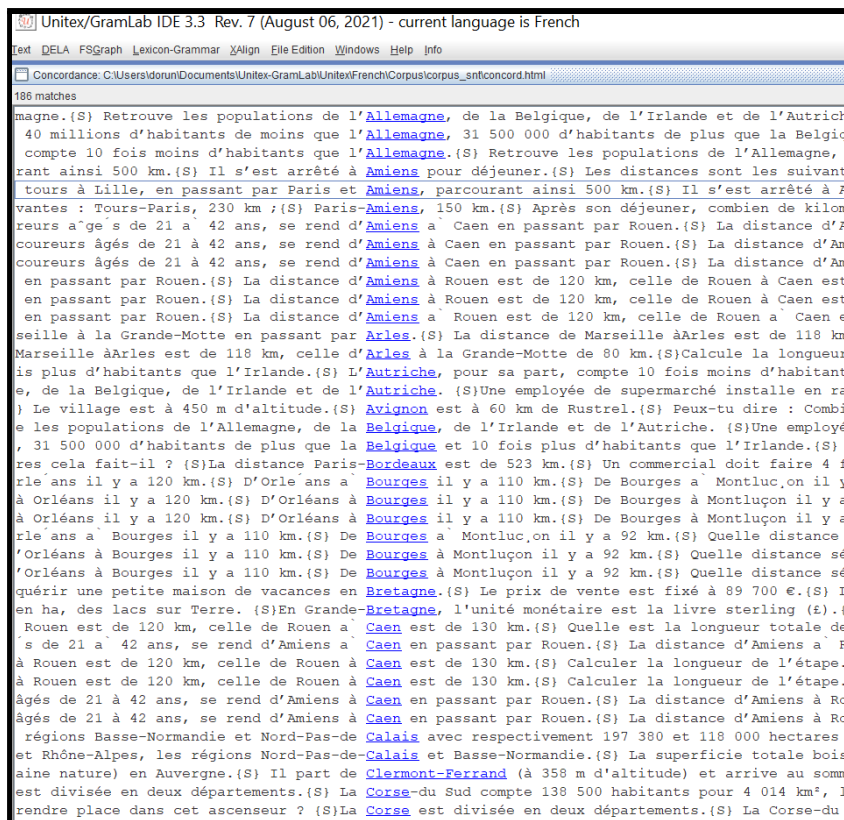
Dans ce cas spécifique, l'automate va reconnaître tous les prénoms communs simples et composés avec un tiret tels que Jean-Luc, par exemple.

Une fois les graphes correspondants à chaque groupe d'entité nommés créés, nous avons décidé de les réunir dans un seul et même graphe (comme pour les expressions temporelles) composé des différents sous-graphes.

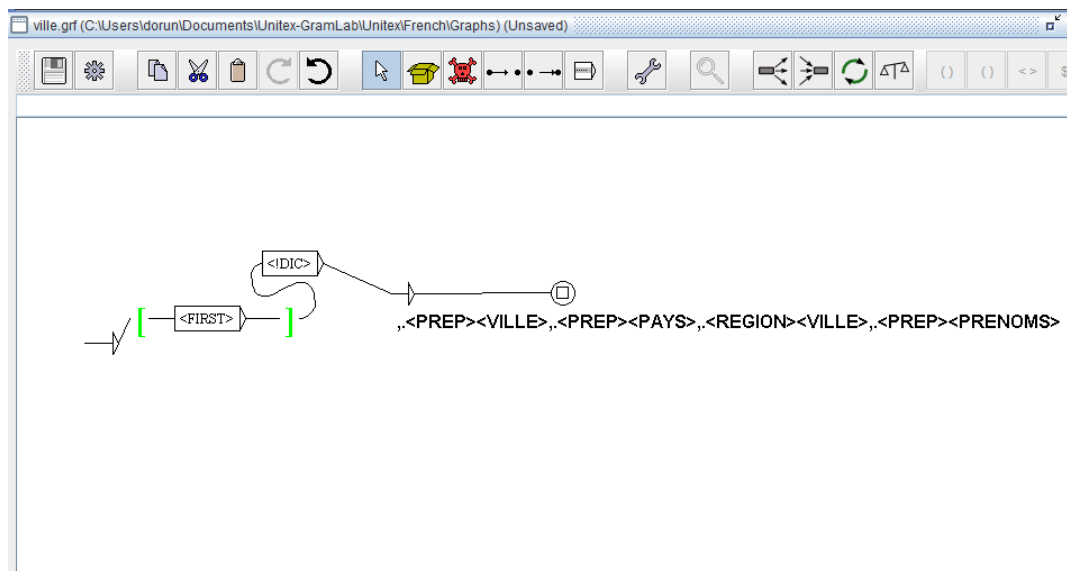
Nous avons procédé de la même manière pour les villes, les pays et les régions. Nous avons constitué les dictionnaires et ensuite nous avons créé un graphe correspondant :



Cela nous a permis de reconnaître un nombre de 186 (villes, pays et régions).



La dernière phase de notre travail consistait à créer un graphe qui puisse reconnaître tous les éléments commençant par une lettre majuscule qui ne sont pas présents dans les dictionnaires : prénoms, pays, régions et villes. Nous avons donc procédé comme suit :



Et voici le résultat obtenu :



Le graphe a trouvé 653 entités nommées ambigus, donc des prénoms étrangers comme Mat, Medhi, ou des villes comme Marrakech, Miami ou encore des régions historiques comme Mésopotamie.

Le résultat est satisfaisant.

Pour conclure la partie qui concerne les entités nommées, nous avons calculé la précision, le rappel et la F-mesure :

	Nb de bons résultats trouvés	Nb de résultats à trouver		Nb de bons résultats trouvés	Nb de résultats trouvés
Précision	653	714	Rappel	653	714
	0,9145658263			0,9145658263	
F-mesure	0,9145658263				

Il semble essentiel de signaler que nos résultats finaux présentent une complication inattendue, comme nous l'a signalé Madame Taravella lors de notre exposé oral.

En effet, étant donné que nous avons pré-traité le texte et supprimé les doublons, nos résultats étaient biaisés, car nous obtenons une précision, un rappel et une F-mesure très élevés. Elle nous a informés qu'il fallait utiliser le corpus de départ sans pré-traitement.

→ Le code python

Concernant la reconnaissance des indices d'opérations, nous nous sommes concentrées sur les termes qui peuvent être ambigus, et qui peuvent parfois induire l'enfant en erreur.

- Voici la liste des termes choisis :

```
addition = ['plus', 'somme']
soustraction = ['différence', 'moins', 'néanmoins']
division = ['séparer', 'sépare', 'chacun', 'chacune', 'par', 'en', 'moyenne']
multiplication = ['produit', 'fois', 'chacun', 'chacune', 'moyenne']
```

Ces termes donnent en général un bon indice quant à l'opération à réaliser, mais ce n'est pas toujours le cas (cf. diapositives de la présentation).

En effet :

- **'plus'** (addition) peut se retrouver dans "n'a plus que" (soustraction)
- **'chacun.e'** (multiplication, division) peut se retrouver dans un problème du type "ils ont chacun tant de billes : combien en ont-ils au total ?" (addition)
- etc...

- Voici le code final :

```
import re # on importe la librairie des regex
import os # on importe la librairie pour manipuler des fichiers

corpus = open("math_corpus.txt", "r") # on ouvre le fichier des énoncés en mode lecture
corpus = corpus.readlines() # on stocke les lignes dans une variable corpus

# on crée nos listes de termes d'indices d'opérations ambigus
addition = ['plus', 'somme']
soustraction = ['différence', 'moins', 'néanmoins']
division = ['séparer', 'sépare', 'chacun', 'chacune', 'par', 'en', 'moyenne']
multiplication = ['produit', 'fois', 'chacun', 'chacune', 'moyenne']

file_resultats = open("resultat_corpus.txt", "w") # on crée un fichier

def traitement(texte): # on crée une fonction traitement
    enonces = set() # on crée un set vide énoncés
    # on incrémente plusieurs compteurs :
    e = 0 # compteur phrases
    nb_add = 0 # compteur addition
    nb_sous = 0 # compteur soustraction
    nb_div = 0 # compteur division
    nb_mult = 0 # compteur multiplication
    # -----
```



```

for phrases in texte: # pour chaque phrase du texte
    e+=1 # on ajoute 1 à e (= nombre d'énoncés, énumération fichier)
    if phrases not in enonces: # si l'énoncé lu n'est pas dans le set d'énoncés
        enonces.add(phrases) # on l'ajoute au set (= permet d'éviter les doublons)
        file_resultats.write(f"\n{e} : {phrases}\n") # et on écrit l'énoncé dans le fichier
        for mot in re.split("\W+",phrases): # pour chaque mot (grâce à la regex) de l'énoncé
            # -----
            if mot in addition: # si le mot est dans la liste addition
                file_resultats.write(f"Mot retrouvé : {mot} --> Addition !\n") # on écrit dans le fichier le
terme trouvé
                nb_add+=1 # et on ajoute 1 au nombre de termes d'addition
            # -----
            if mot in soustraction: # si le mot est dans la liste soustraction
                file_resultats.write(f"Mot retrouvé : {mot} --> Soustraction !\n") # on écrit dans le
fichier le terme trouvé
                nb_sous+=1 # et on ajoute 1 au nombre de termes de soustraction
            # -----
            if mot in division: # si le mot est dans la liste division
                file_resultats.write(f"Mot retrouvé : {mot} --> Division !\n") # on écrit dans le fichier le
terme trouvé
                nb_div+=1 # et on ajoute 1 au nombre de termes de division
            # -----
            if mot in multiplication: # si le mot est dans la liste multiplication
                file_resultats.write(f"Mot retrouvé : {mot} --> Multiplication !\n") # on écrit dans le
fichier le terme trouvé
                nb_mult+=1 # et on ajoute 1 au nombre de termes de multiplication
        else: # si la phrase est déjà dans le set
            break # on arrête
# et on affiche nos compteurs
print('Termes "addition" : ',nb_add)
print('Termes "soustraction" : ',nb_sous)
print('Termes "division" : ',nb_div)
print('Termes "multiplication" : ',nb_mult)
print(len(enonces))

traitement(corpus) # on appelle notre fonction sur le corpus
file_resultats.close() # on ferme le fichier de resultats créé

```

➔ Problèmes rencontrés

Le corpus présentait des doublons, mais nous avons réussi à les éliminer avec une simple fonctionnalité de Google Sheets. Cependant, certains doublons semblaient persister même

après. Le problème venait du fait qu'il y avait parfois simplement un espace en plus ou en moins, et par conséquent, la chaîne de caractère entière était traitée comme étant unique.

Le corpus présentait également des fautes d'orthographe : ce problème s'est avéré en être un lorsque nous sommes passées sur Unitex. En effet, certaines expressions temporelles ou entités nommées ne pouvaient pas être reconnues parce qu'elles étaient mal écrites, ou collées au mot précédent/suivant, ou parce que l'apostrophe n'était pas reconnue (problème typographique), etc...

→ Améliorations possibles

Nous aurions pu nous attarder sur les doublons persistants en nous occupant des espaces en trop ou en moins, pour vraiment tous les éliminer.

Sur Unitex, nous aurions pu affiner encore plus nos graphes pour qu'ils reconnaissent les expressions temporelles et entités nommées dans leur entièreté (seule une partie est parfois prise en compte).

Autre amélioration : au début, nous sommes parties sur l'idée de reconnaître les opérations mathématiques à réaliser pour chaque énoncé en repérant des sortes de "patterns" et en nous basant sur des termes spécifiques et distinctifs (ex : les termes "dépenser", "enlever", ou pattern "n'a plus que"... indiquent qu'il faut soustraire).

OBJECTIF : repérer et définir des patterns, des mots-clés qui indiquent quelles opérations il faut faire

- sur Unitex : penser au principe des boîtes

- sur python : création de listes (addition, soustraction, division, multiplication...) avec regex pour repérer les patterns

Adv. temporel	Situation	Addition (+)	Soustraction (-)	Division (÷)	Multiplication (x)	Adv. temporel	Finalité	Note
	avait		dépensé, acheter			maintenant	somme	passé / présent (SITUATION / FINALITÉ : montre qu'on ne peut pas se fier au temps des verbes pour rep
	a	donne				maintenant	a donné	présent / passé (SITUATION / FINALITÉ : montre qu'on ne peut pas se fier au temps des verbes pour rep
	a	donne				maintenant	avait	présent / passé (SITUATION / FINALITÉ : montre qu'on ne peut pas se fier au temps des verbes pour rep
	a		enlève			maintenant	a	présent / présent (SITUATION / FINALITÉ : montre qu'on ne peut pas se fier au temps des verbes pour re
	a		enlève, n'a plus que			maintenant	a	
	pèse	de plus que						
	pèse		de moins que					
	a acheté... à...						montant	--> "montant" induit addition
	(il y a...) NUM... NUM...						en tout / total(e)	--> "en tout" induit addition
	de... à... il y a...							pas d'indices d'opération ! repose sur l'énumération = ADDITION
	composer<V>... NUM... et... NUM...						composé	pas d'indices d'opération ! repose sur l'énumération = ADDITION
en une semaine	ont consommé... NUM pour... NUM pour...						en tout	--> "en tout" induit addition + énumération
	11 km par jour				en une=1 semaine, en NUM jours, en NUM jours			
	NUM fois							
	NUM... NUM... NUM					en février, en mars ,		pendant les trois automatiquement on va devoir reconnaître trois en mot et en chiffre
					valeur de NUM, le solde, chaque mensualité			

→ Sources

- ❖ Les communes les plus peuplées de France. L'internaute [en ligne]. Disponible sur <<https://www.linternaute.com/>> [consulté le 29/03/2022]
- ❖ Anne-laure Mignon, 23 mars 2022. Le Figaro [en ligne]. Disponible sur <<https://www.lefigaro.fr/>> [consulté le 29/03/2022]

- ❖ Cartes des départements de France. Cartes de France [en ligne]. Disponible sur <<http://www.cartesfrance.fr/>> [consulté le 29/03/2022]
- ❖ Liste des prénoms féminins français. Prénom [en ligne]. Disponible sur <<https://prenom.org/>> [consulté le 29/03/2022]
- ❖ Liste des prénoms masculins français. Prénom [en ligne]. Disponible sur <<https://prenom.org/>> [consulté le 29/03/2022]