La grande aventure du traitement automatique des langues

Scénario : Carlos Ramisch et Marion Cluzel
Dessins : Marion Cluzel

La grande aventure du Traitement automatique des langues

Scénario : Carlos Ramisch et Marion Cluzel
Dessins : Marion Cluzel

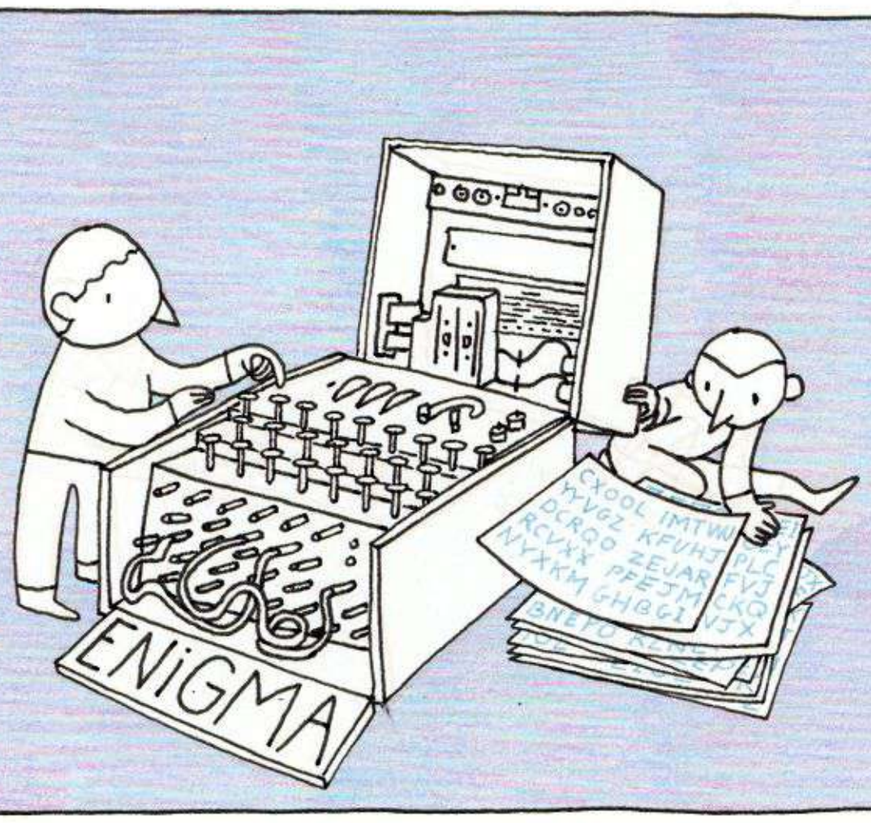
Le traitement automatique des langues (TAL)
est la partie de l'informatique
qui s'intéresse au langage humain,
et qui développe des outils pour nous aider
à écrire, lire, communiquer
(comme chatGPT ou Google Translate).

Ces dernières années, les technologies
du langage sont devenues très présentes
dans nos smartphones, enceintes connectées,
automobiles, moteurs de recherche, etc.

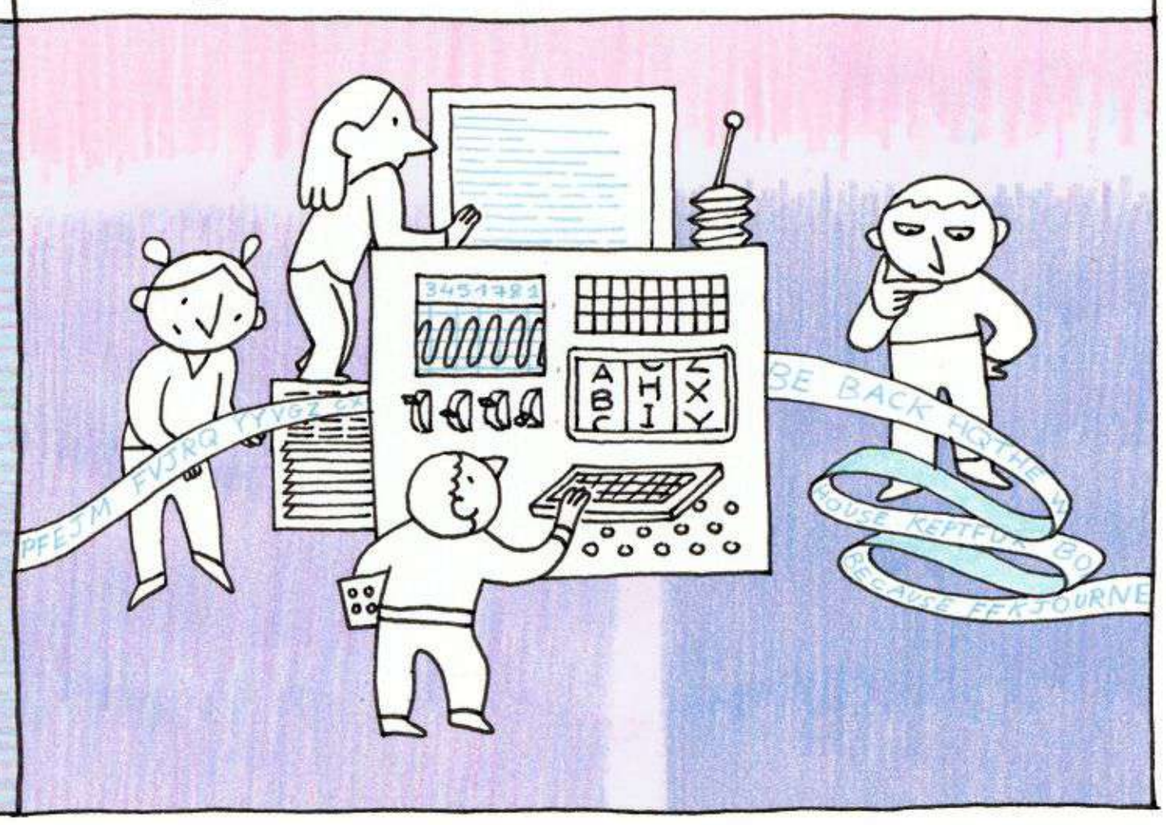
Comment ça marche ?
Pourquoi, parfois, ça ne marche pas ?
Peut-on leur faire vraiment confiance ?

Pour comprendre l'histoire du TAL, il faut revenir aux débuts de l'informatique.

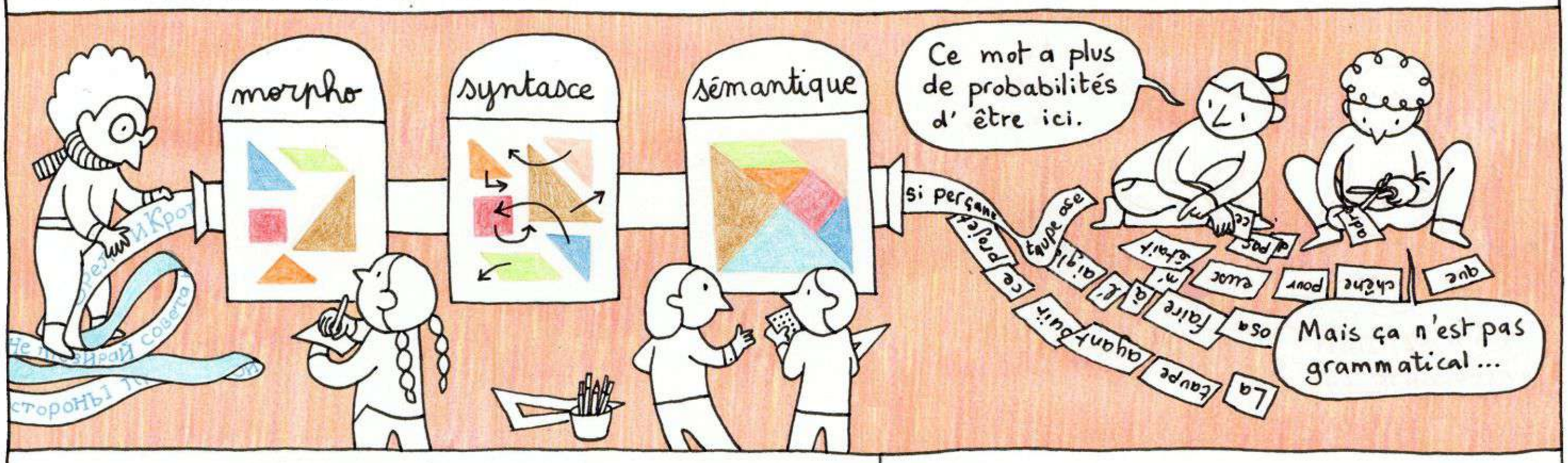
Pendant la Seconde Guerre mondiale, les Allemands ont inventé ENIGMA, une machine pour crypter leurs messages.



Les Anglais ont alors créé le tout premier ordinateur, capable de décrypter les messages d'ENIGMA.



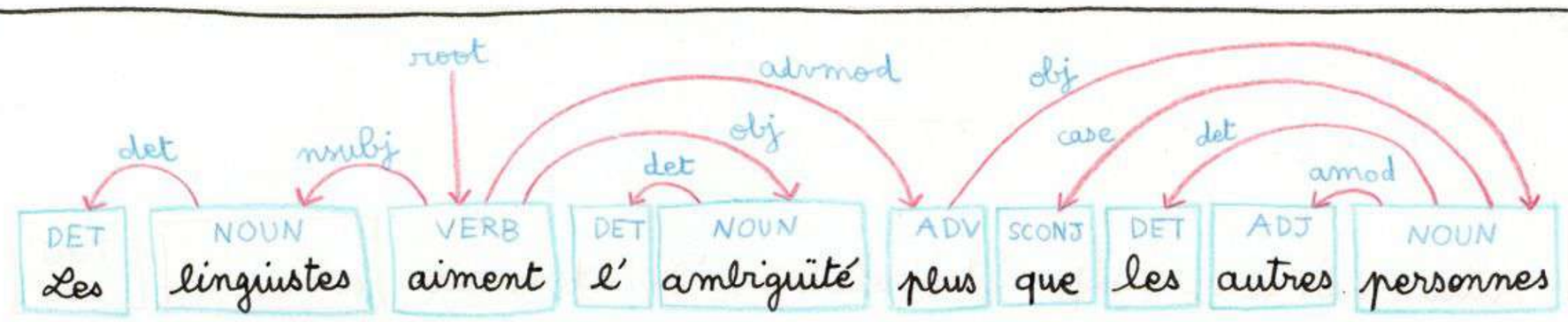
Les États-Unis vont ensuite s'inspirer de ce succès pour développer, pendant la guerre Froide, des machines pour "décrypter" le russe en le traduisant vers l'anglais : c'est le premier système de TAL !



Dans ces premiers systèmes, on représente la morphologie, la syntaxe et la sémantique par des règles complexes : c'est la MÉTHODE EXPERTE.

Cela demande beaucoup d'expert.es en linguistique et en informatique, et ne marche pas toujours très bien.

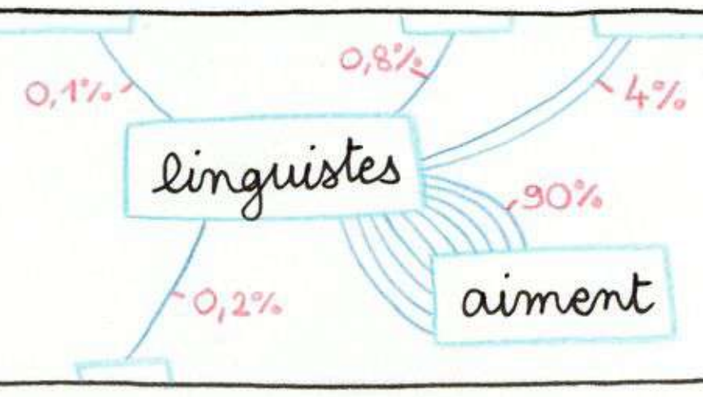
Les années 1990 voient le développement rapide d'internet : une source presque infinie de DONNÉES. Désormais, au lieu d'écrire des règles, les linguistes annotent ces données.



relation linguistique

On peut utiliser des statistiques (mathématiques) sur ces données d'entraînement pour estimer les probabilités de chaque relation linguistique.
Une fois l'ordinateur "entraîné", il prédit ces relations tout seul.

C'est le début du MACHINE LEARNING ou APPRENTISSAGE AUTOMATIQUE.

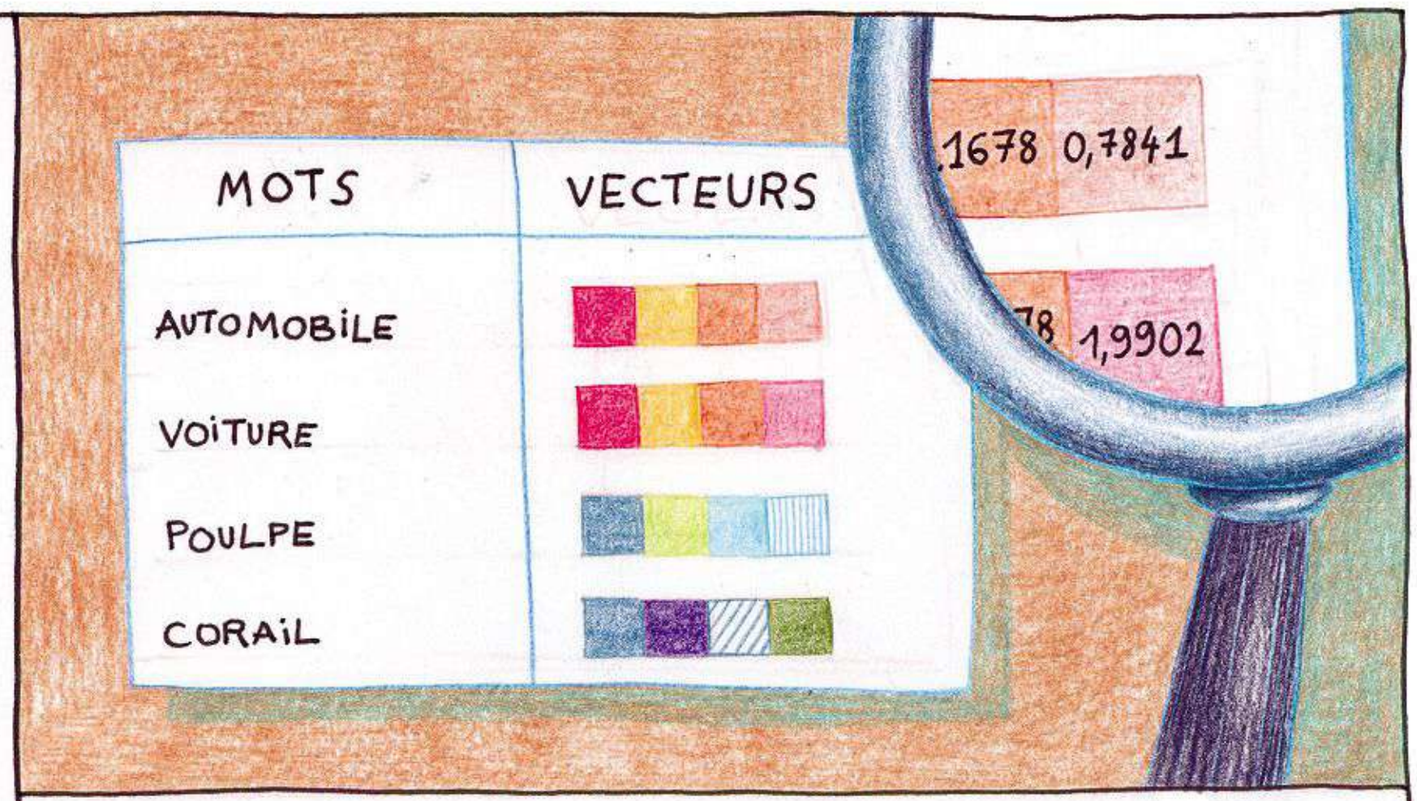


[LINGUISTES] ici a 90 % de chances d'être le sujet de [AIMENT].

Au début des années 2010, un nouveau type de machine learning prend de l'essor : le DEEP LEARNING.

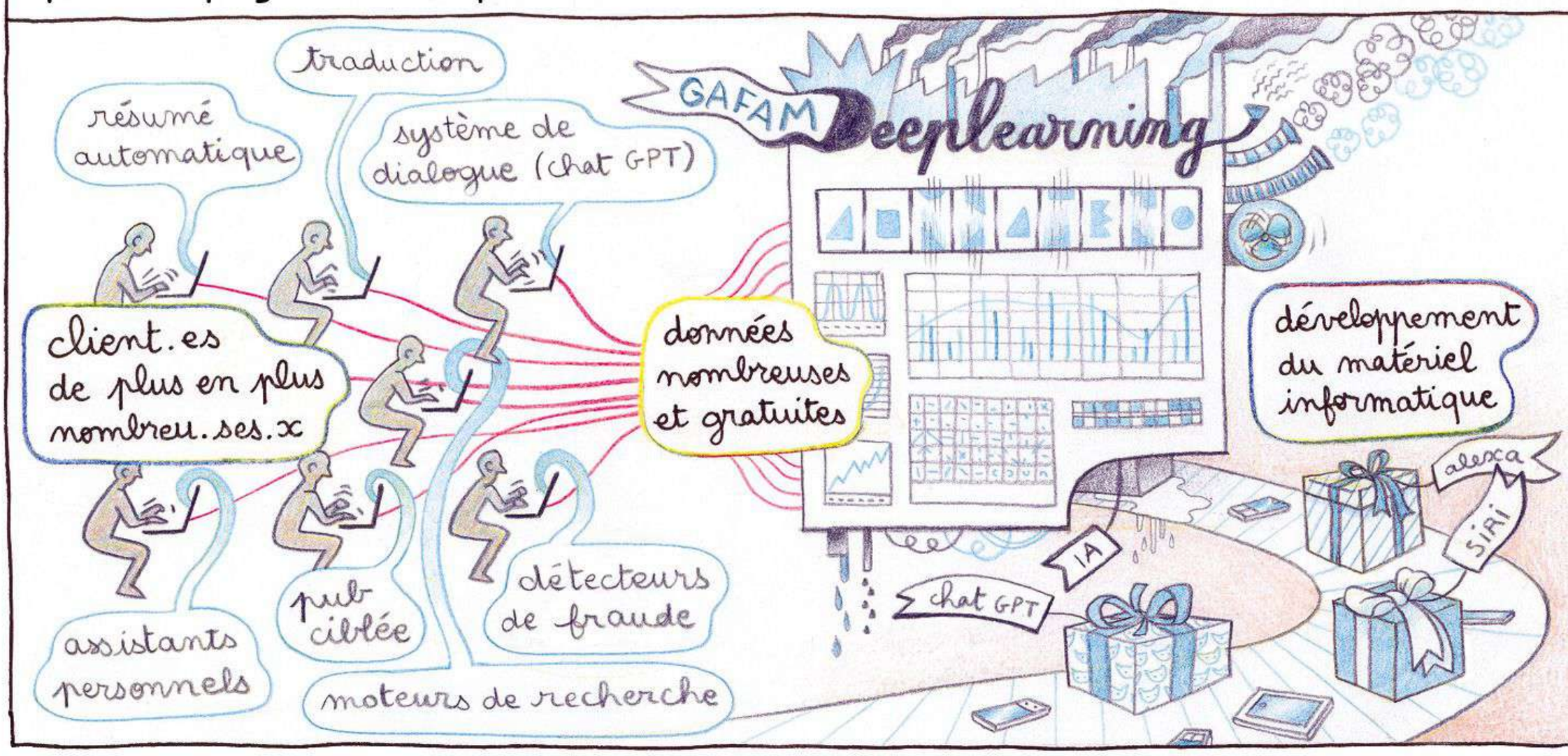
L'idée est de transformer chaque mot en une suite de nombres, appelée VECTEUR.

Cela permet de représenter mathématiquement les similarités entre les sens des mots, et de traiter beaucoup plus de données.



Désormais, les calculs et équations jouent le rôle qui était assuré auparavant par des traitements linguistiques spécifiques. Et ça marche beaucoup mieux !

Le deep learning suscite un enthousiasme fou et des résultats impressionnants.
Il améliore les performances et démocratise des technologies qui accompagnent notre quotidien:



Financée en grande partie par les GAFAM (multinationales du numérique), cette technologie s'intègre dans leurs produits, et envahit notre quotidien...

Cependant, les systèmes à base de deep learning ont quelques limitations, dont voici les deux principales.

Le manque de diversité

Par exemple, un système de questions-réponses semble imbattable sur des questions "connues", qui ressemblent aux données d'entraînement du système.

Oh ! Une coccinelle !
Regarde, elle se pose sur la fraise.
Et maintenant, elle la mange.

Que mange la coccinelle ?

La coccinelle mange la fraise.

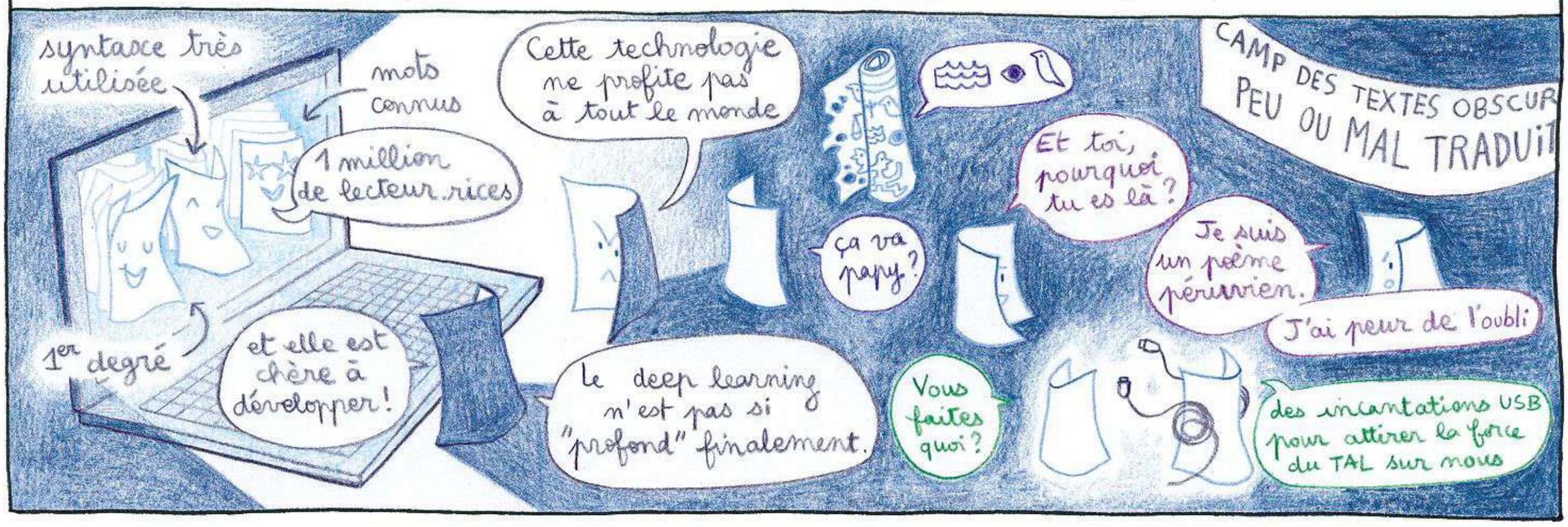
J'ai longtemps hésité à écrire un livre sur la femme. Le sujet est irritant mes ; et il n'est pas neuf. La querelle du féminisme a fait couler assez est à peu près close : n'en parlons plus. On en parle encore cependant. Et les volumineuses sottises débitées pendant ce dernier siècle aient bea blème. D'ailleurs y a-t-il un problème ? Et quel est-il ? Y a-t-il même des éorie de l'éternel féminin compte encore des adeptes ; ils chuchotent : « restent bien femmes » ; mais d'autres gens bien informés – et l quefois – soupirent : « La femme se perd, la femme est perdue. » On ne te encore des femmes, s'il en existera toujours, s'il faut ou non le souha s occupent en ce monde, quelle place elles devraient y occuper. « Où sor andait récemment un magazine intermittent[1]. Mais d'abord : qu'est-ce *ta mulier in utero* : c'est une matrice », dit l'un. Cependant n de c

Qu'est-ce qu'une femme ?

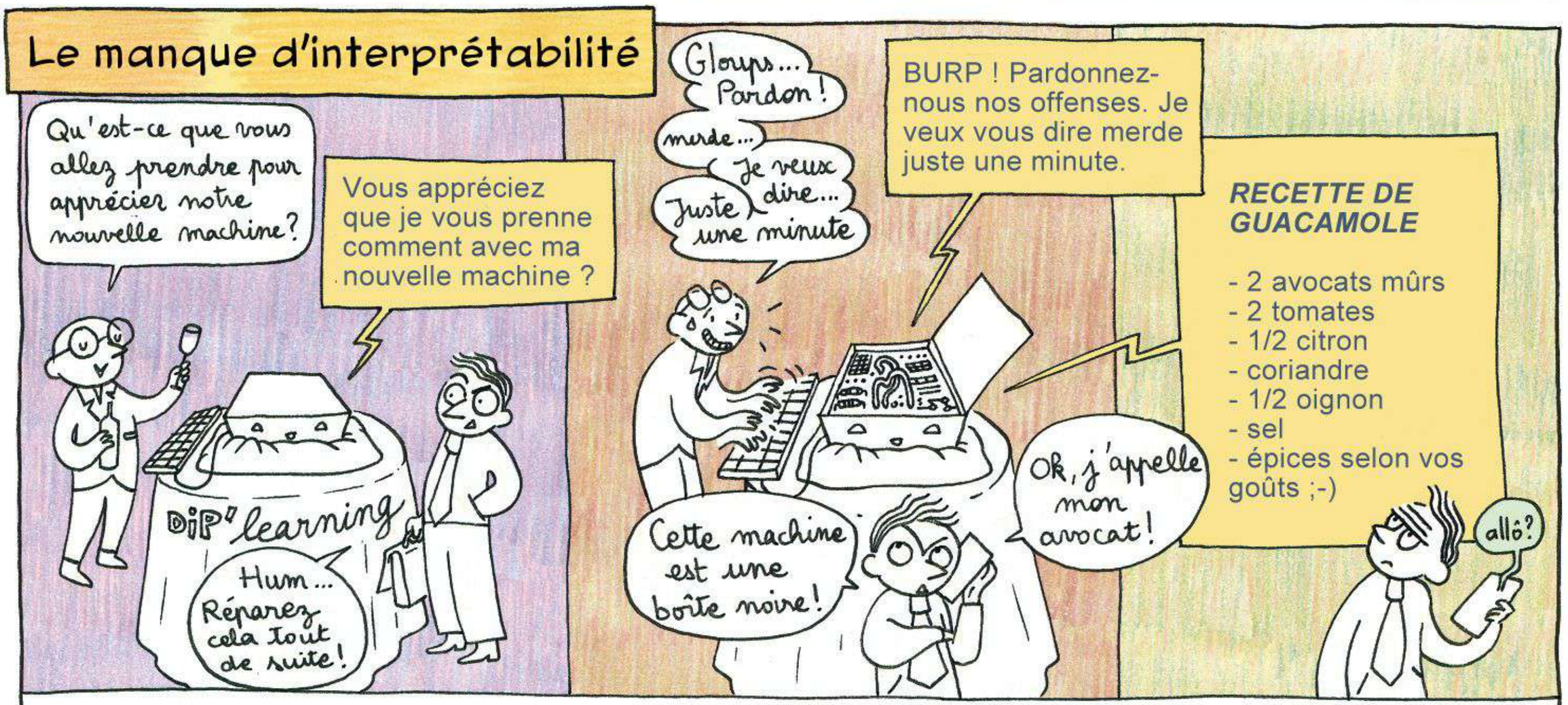
La femme est perdue.

Mais dès que la question est en dehors de sa zone de confort, l'ordinateur n'y arrive plus.

Autrement dit, ces systèmes ne sont pas très ROBUSTES par rapport aux données qu'ils ont utilisées pour apprendre les statistiques, et ne rendent pas compte de toute la DIVERSITÉ propre aux langues.



Le manque d'interprétabilité



Lorsqu'une erreur apparaît, on ne peut pas comprendre d'où elle vient, ni la corriger facilement, car les formules mathématiques sont très OPAQUES.

Houston, on a un problème. Le deep learning n'est pas assez fiable.

Comment ça ?

L'intelligence de la machine est limitée aux textes d'entraînement qui lui ont été donnés.

Autrement dit, TOUT CE QUI TRAÎNE SUR INTERNET !!!

Et je n'ai que faire de recettes de cuisine pour les problématiques de nos client.es !

Et en plus, elle est complètement chiffrée, ce qui la rend très opaque pour agir directement dessus.

Je vois. J'ai une idée.

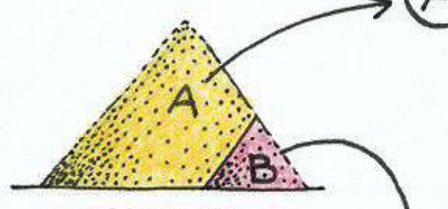
Mettez un casque et écoutez ma voix.

Hein ?!

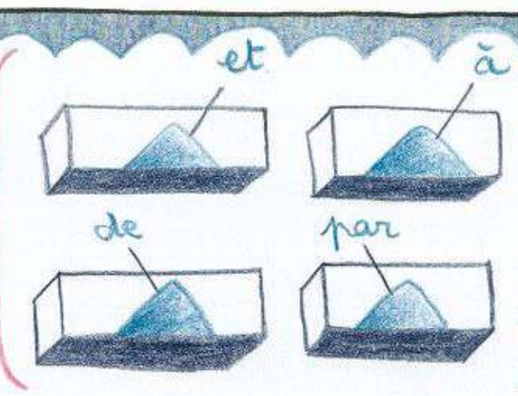
Vous verrez, c'est plus immersif.

Bien sûr, on pourrait embaucher PLUS de personnes pour engranger ENCORE PLUS de données.

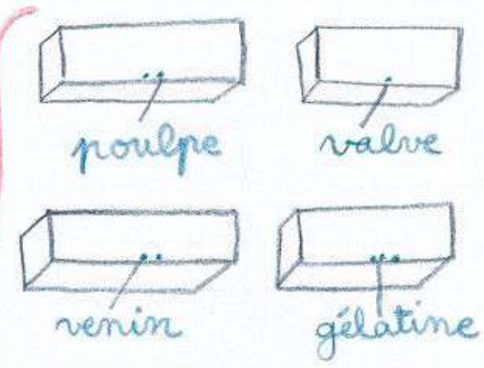
Imaginez qu'un texte est un tas de sable et que chaque grain qui le compose est un mot.

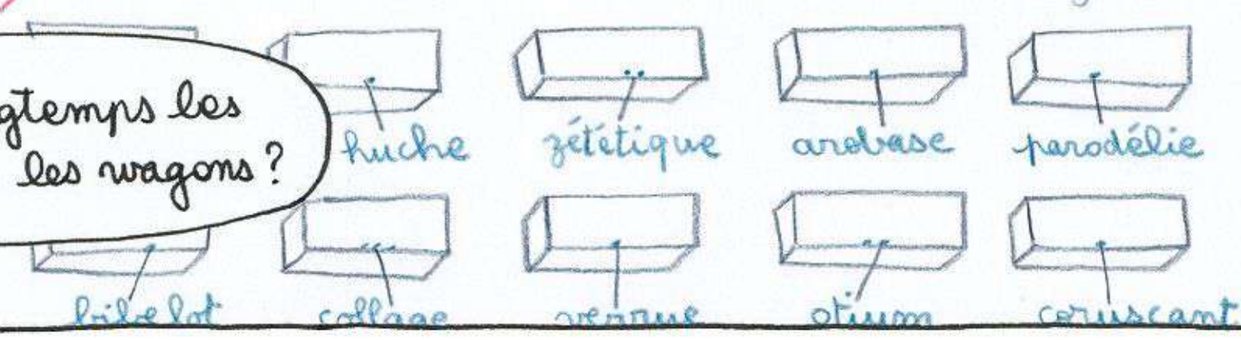


(A) Ici, les mots qui apparaissent très souvent et sont bien connus remplissent chacun un wagon qui les contient



(B) Puis les mots qui apparaissent en grande variété mais en petite quantité, remplissent d'autres wagons... toujours plus nombreux...





Ça va durer longtemps les petits grains dans les wagons ?

Les nouveaux mots sont présents en grande variété et apparaissent de moins en moins dans les textes.

C'est ce qu'on appelle la loi de ZIPF.

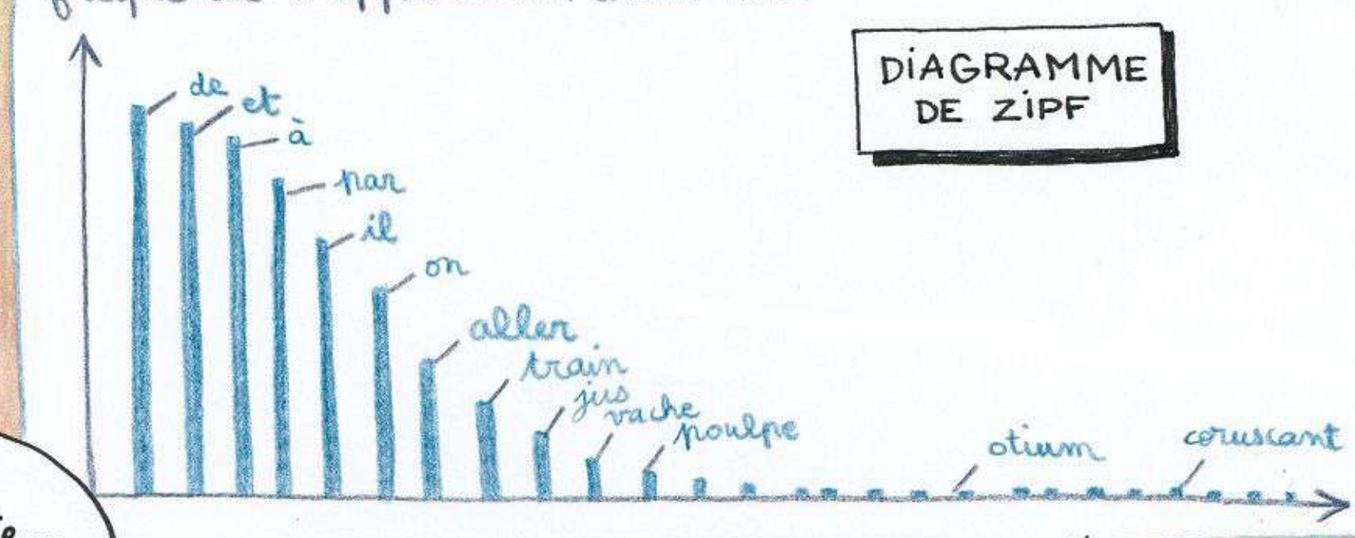


DIAGRAMME DE ZIPF

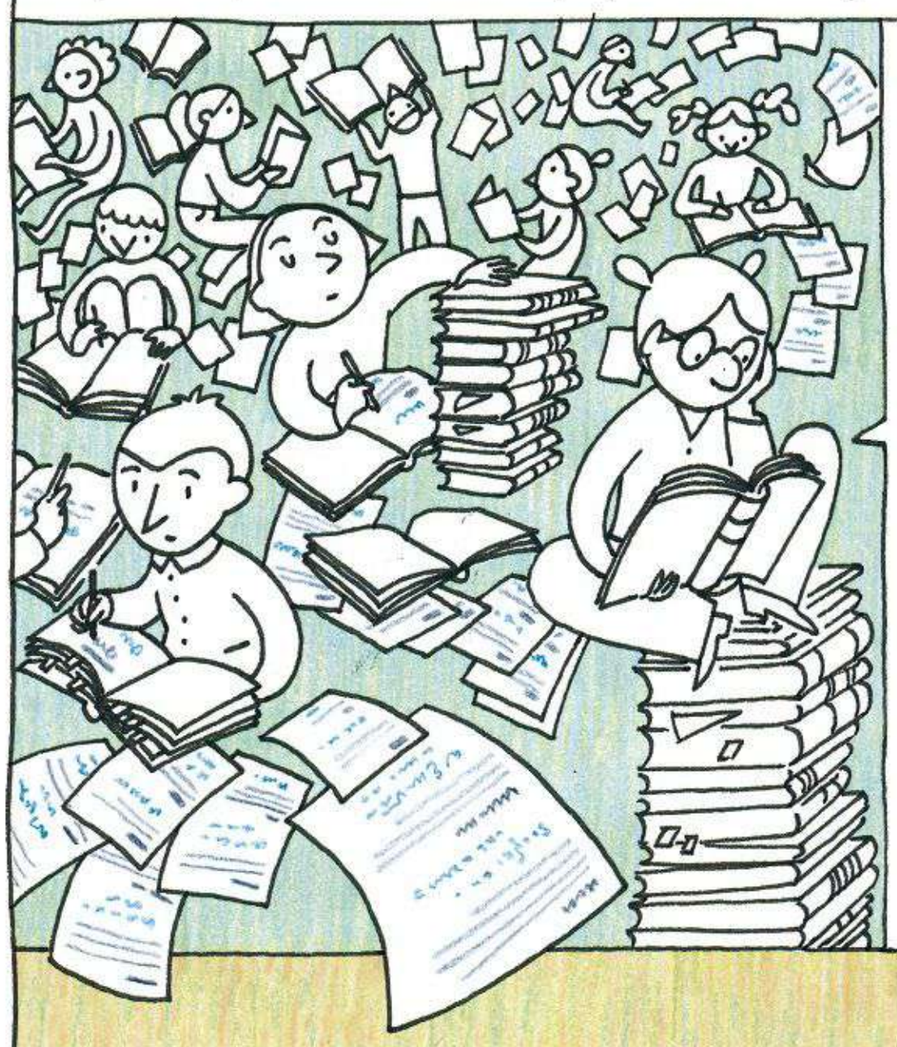
Mais c'est de la folie... Tant d'efforts déployés pour trouver un nouveau mot qui sera si peu utilisé...

Il doit bien y avoir une autre option que de compter les moutons !

les wagons.

oui, oui, les wagons.

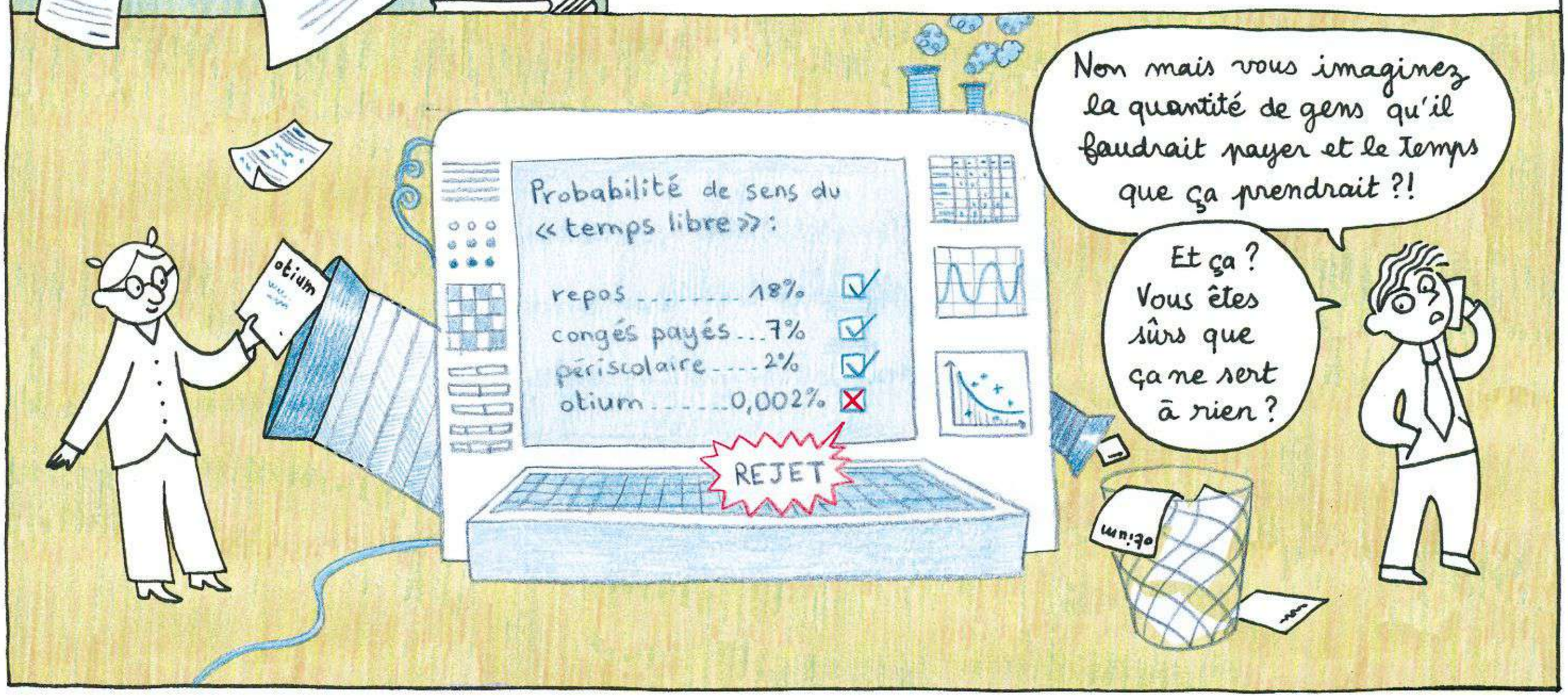
Alors oui, alternativement, on pourrait développer à la main des lexiques (dictionnaires pour ordinateur) très précis qui décrivent en détail tous les sens possibles pour chaque mot.



L'otium

L'otium est un terme latin remontant au milieu du IIe siècle av. J.-C. qui recouvre différentes formes et significations dans le champ du temps libre. C'est le temps durant lequel une personne profite du repos pour s'adonner à la méditation, au loisir studieux. C'est aussi le temps de la retraite à l'issue d'une carrière publique ou privée, par opposition à la vie active, à la vie publique.

C'est un temps, sporadique ou prolongé, de loisir personnel aux implications intellectuelles, vertueuses ou morales avec l'idée d'éloignement du quotidien, des affaires (negotium = neg-otium), et d'engagement dans des activités valorisant le développement artistique ou intellectuel (éloquence, écriture, philosophie). L'otium revêt une valeur particulière pour les hommes d'affaires, les diplomates, les philosophes ou les poètes.



Le projet SELEXINI propose une approche à mi-chemin entre le deep learning et les dictionnaires : on s'inspire du deep learning pour découvrir automatiquement, à partir des textes, des informations sur les SENS des mots, et ainsi rendre les systèmes plus compréhensibles pour les humains.

LA PRÉPARATION

On collecte un grand corpus de textes sur le web. Tels qu'ils y figurent, les mots ne sont pas pratiques à manipuler. Leur préparation consiste à :

A) distinguer les catégories des mots selon leur contexte

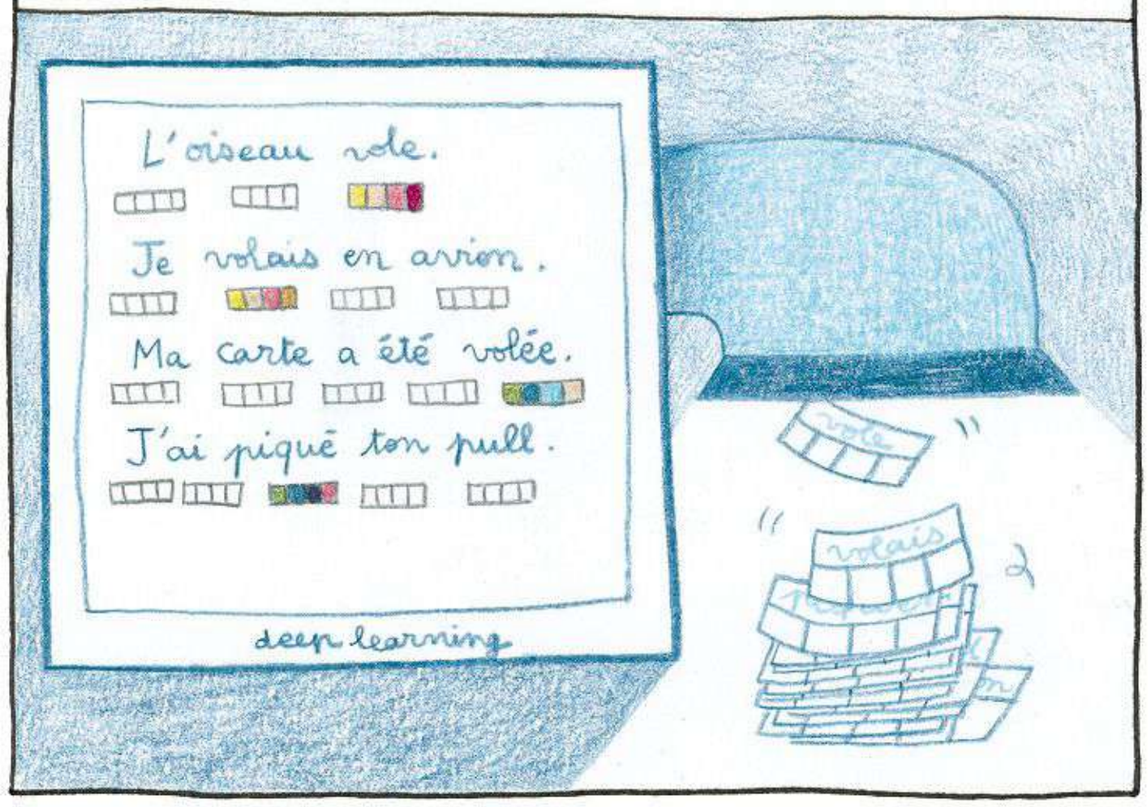
B) regrouper les différentes formes fléchies de chaque mot

C) identifier les expressions idiomatiques, dont le sens ne se décompose pas

LA CRÉATION DU LEXIQUE

On utilise les mots et leurs contextes d'apparition dans le grand corpus pour faire des clusters (des regroupements, pas des confinements hein).

Un modèle de deep learning est utilisé pour générer des vecteurs pour chaque mot qui apparaît dans le grand corpus.



Ces vecteurs peuvent être comparés mathématiquement, et cela nous indique à quel point leurs sens sont proches.

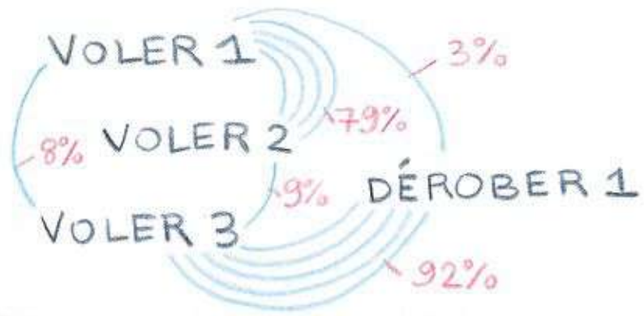
Cette comparaison permet de séparer dans des clusters différents des mots qui ont la même forme mais des sens différents (polysémiques), et de regrouper dans le même cluster des mots dont les formes n'ont rien à voir, mais qui veulent dire la même chose (synonymes).

deeplearning

L'oiseau vole. → VOLER 1
Je volais en avion. → VOLER 2
J'ai volé ton idée. → VOLER 3
Ma carte à été dérobée. → DÉROBER 1
Je te pique ta veste. → PIQUER 1
L'avion

COSINUS

	VOLER 2	VOLER 3	DÉROBER 1
VOLER 1	79%	8%	3%
VOLER 2	100%	9%	4%
VOLER 3	9%	100%	92%



LEXIQUE AUTOMATIQUE

Cluster 1 :
VOLER 1 ; VOLER 2 ; PLANER 1

Cluster 2 :
VOLER 3 ; DÉROBER 1 ; PIQUER

Autrement dit, le lexique contient des clusters trouvés de maniere 100 % automatique, à partir des similarités entre vecteurs de mots.

L'APPLICATION Ce lexique construit automatiquement peut contribuer à une meilleure interprétabilité et diversité dans les systèmes de TAL.

De la même façon qu'un dictionnaire peut nous aider à comprendre un mot dans un texte, l'ordinateur utilise le lexique sémantique pour prendre ses propres décisions.

Ma carte a été dérobée.

Qu'est-ce qui a été dérobé ?

et "dérober", ça veut dire quoi ?

La carte

dérober = voler (92 %)

regarde lulu

Cluster 2 :
VOLER 3 ; DÉROBER 1
PIQUER 1 ; PRENDRE

Mais comment il peut le savoir l'ordinateur ?!

J'ai vu le « le tableau a été volé » dans mes données d'entraînement, et le cluster 2 m'informe que « voler » et « dérober » sont synonymes.

Notre espoir est de mieux tenir compte de la diversité linguistique dans les systèmes de TAL, tout en sachant interpréter leurs décisions à l'aide des informations contenues dans le lexique.

Ajouter des infos qui aident les humains à interpréter les mathématiques derrière les modèles nous redonne du pouvoir pour mieux maîtriser cette technologie, comprendre pourquoi ça marche, ou ne marche pas.

Auteur.rices : Carlos Ramisch et Marion Cluzel

Relecteur.rices : Marie Candito, Estelle Dehame, Laure Dupont, Benoit Favre, Amandine Goy, Charlotte Rousselle, Agata Savary, Manon Scholivet, Camille Veillard

Correctrice : Laure Dupont

Soutiens : Laboratoire d'Informatique et des Systèmes (LIS), Campus France - PHC Ulysses, Aix Marseille Université, Agence Nationale de la Recherche via le projet SELEXINI (ANR-21-CE23-0033-01).

ambiguïté
$f := a+b \left\{ \frac{x}{2} \sqrt{z} \right.$