Université de Franche-Comté

**Appariement phrastique pour un système de questions-réponses basé sur une FAQ dans le domaine de la santé**

Par Ana Calatayud

Master Langues, Littératures et Civilisations Etrangères et Régionales (LLCER)

Parcours Traitement Automatique des Langues (TAL)

Faculté des Sciences du Langage, de l’Homme et de la Société

Directrice de mémoire : Dr. Iana Atanassova

Mémoire présenté à la Faculté des Sciences du Langage, de l’Homme et de la Société

25 mai 2018



Remerciements

**A mes professeures, Iana et Izabella**

Merci à Izabella Thomas et à Iana Atanassova pour leur enseignement. Un merci tout particulier au Dr. Atanassova d’avoir accepté de diriger ce mémoire et de m’avoir guidé, conseillé et encouragé.

**A mes parents, Christophe et Carole**

Votre soutien a été un moteur dans la réalisation de ce mémoire et j’espère vous rendre fiers. Sachez que ma gratitude et mon amour sont inconditionnels.

**A ma sœur, Aude**

Tu as toujours été présente, tu m’as toujours soutenu et encouragé. Sache que quand ton tour viendra ça sera réciproque. Merci à toi d’être une sœur géniale.

**A toi Léo**

Merci de m’avoir rassuré dans les moments de doute, de m’avoir poussé vers le haut quand j’avais envie de baisser les bras, merci pour tout.

**A mes camarades, Yağmur, Elodie et Salah**

Bien plus que des camarades de classe vous êtes devenus des amis. Merci à vous d’avoir été là, merci de m’avoir soutenu. Ce mémoire, je le dédis à la promotion soudée que nous sommes.

**A mon équipe de stage (Houda, Dialekti, Abderrahim)**

Table des matières

[I. Introduction 5](#_Toc5816010)

[1. Les systèmes question-réponse 5](#_Toc5816011)

[2. Les Foires Aux Questions 7](#_Toc5816012)

[3. Problématique et objectifs. 7](#_Toc5816013)

[4. Etat de l’art : 8](#_Toc5816014)

[1) SQR dans le domaine de la santé. 8](#_Toc5816015)

[2) Les systèmes QR basés sur des FAQ 11](#_Toc5816016)

[5. Concepts clés du sujet 13](#_Toc5816017)

[II. Méthodologie pour l’appariement de questions dans un système QR basé sur une FAQ 15](#_Toc5816018)

[6. Hypothèses de recherche 15](#_Toc5816019)

[7. Domaine 17](#_Toc5816020)

[8. Ressources 17](#_Toc5816021)

[3) Corpus de questions 17](#_Toc5816022)

[4) Corpus de réponse 19](#_Toc5816023)

[9. Types de questions 19](#_Toc5816024)

[1) Les questions fermées. 19](#_Toc5816025)

[2) Les questions ouvertes 21](#_Toc5816026)

[III. Théories linguistiques latentes 21](#_Toc5816027)

[10. La Semantic Theory 21](#_Toc5816028)

[11. Les limites de la syntaxe 23](#_Toc5816029)

[12. La paraphrase et la Théorie Sens-Texte 24](#_Toc5816030)

[3) Fondements de la Théorie Sens-Texte 24](#_Toc5816031)

[4) Les composantes du Modèle Sens-Texte 26](#_Toc5816032)

[13. La paraphrase en TAL 28](#_Toc5816033)

[14. Le distributionnalisme 28](#_Toc5816034)

[IV. Implémentation du système 31](#_Toc5816035)

[15. Première modélisation 31](#_Toc5816036)

[1) Présentation de NLTK et TreeTagger 31](#_Toc5816037)

[2) Segmentation. 31](#_Toc5816038)

[3) Lemmatisation. 32](#_Toc5816039)

[16. Vectorisation lexicale et phrastique 36](#_Toc5816040)

[a) La représentation en bag-of-words (BOW) 36](#_Toc5816041)

[17. Les différentes mesures de similarité 40](#_Toc5816042)

[4) Méthodes de similarité textuelle. 40](#_Toc5816043)

[b) Distance de Jaccard 40](#_Toc5816044)

[c) Similarité cosinus 41](#_Toc5816045)

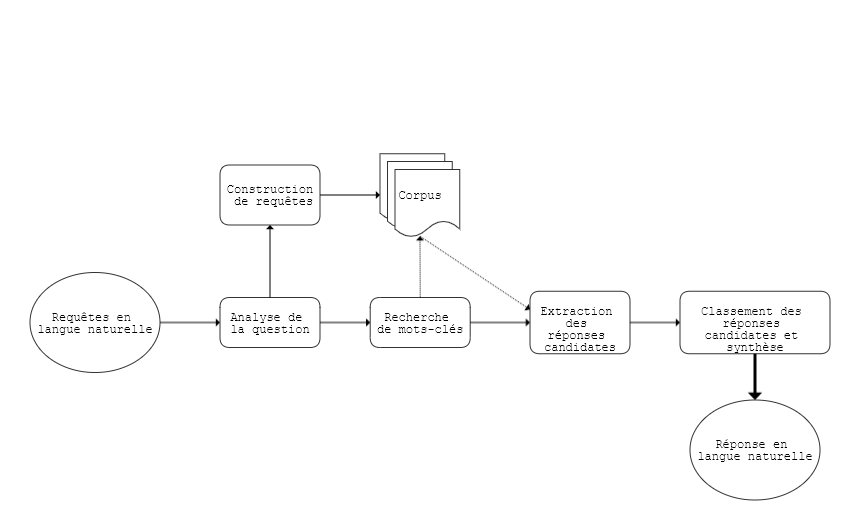
[5) Mesures de similarité sémantique 42](#_Toc5816046)

[6) Analyse des questions 42](#_Toc5816047)

[V. Annexes et bibliographie 43](#_Toc5816048)

# Introduction

## Les systèmes question-réponse

De nos jours, l’accès à l’information est rendu aisé grâce à internet. On peut cependant aussi l’envisager comme un ‘cadeau empoisonné’ dans la mesure où la quantité de données qui y transitent est titanesque, rendant leur exploitation efficace impossible. Les moteurs de recherche tels que Google ou Yahoo ne facilitent pas la tâche puisqu’une simple requête engendre des milliers de résultats. L’utilisateur doit alors perdre un temps considérable à chercher la réponse à sa requête dans cet amas de données hétérogènes. C’est dans ce contexte que s’inscrivent les systèmes de question réponse (SQR). Un système de question réponse permet à un utilisateur de formuler une requête en langue naturelle. Cette requête est ensuite analysée, puis des réponses potentielles sont recherchées dans un corpus composé en général de données et de supports hétérogènes. La réponse la plus pertinente est ensuite extraite puis retournée à l’utilisateur, le plus souvent sous forme de phrase formulée en langue naturelle.

*Figure 1. Architecture basique d’un SQR.*

Les premiers systèmes de question-réponse, BASEBALL (Green et al. 1961), programmé pour répondre à des questions concernant le champion de baseball américain d’une certaine année, et LUNAR (Woods), programmé pour répondre à des questions concernant la lune, furent développés autour des années 1960-1970. Ces permiers systèmes possédaient cependant de nombreuses restrictions sur la manière d’utiliser la langue, mettant ainsi en question leur qualification de SQR puisque les requêtes doivent être formulées en langue naturelle. Les SQR font un véritable bond en avant grâce aux conférences TREC (*Text REtrival Conference*) qui ont débuté en 1992 et qui s’intéressent à l’extraction d’information, un élément essentiel des SQR. Ce n’est cependant qu’en 1999 que ces conférences ont commencé à intégrer des campagnes d’évaluation de SQR, grâce auxquelles de nombreux systèmes comme QRISTAL (Laurent and Séguéla 2004), QALC (Ligozat 2004), FIDGI (Moriceau and Tannier 2009) ou encore PiQASso (Antonio et al. 2001) ont vu le jour. Ces systèmes sont à domaine ouvert (*open-domain*), c’est-à-dire qu’ils peuvent répondre à n’importe quelle question générique, comme la date de naissance de Shakespeare ou bien la capitale de l’Ouzbékistan.

Il existe des systèmes question-réponse à domaine restreint (*domain-dependent*). L’un des domaines encore peu pourvu en matière de SQR est celui de la santé et du médical. L’une des raisons qui pourrait expliquer ce vide est que le domaine de la santé est un domaine sensible dans le sens où les requêtes formulées doivent obtenir des réponses précises, non ambigües et surtout correctes. Par exemple, avec un système QR qui ne ferait qu’une analyse superficielle et rudimentaire des données textuelles, une requête telle que « qu’est-ce que l’anémie ? » recherchée dans un corpus qui comporterait la phrase « […] causé par le manque de magnésium ou l’anémie » obtiendra potentiellement comme réponse « un manque de magnésium », ce qui n’est pas correct et ce qui pourrait induire l’utilisateur en erreur. Une autre raison pourrait être celle de la difficulté de constituer une base de données suffisante et fiable dans un domaine restreint . Depuis le début des années 2000 cependant, quelques SQR dans le domaine de la santé ont vu le jour.

## Les Foires Aux Questions

Aussi appelées *Frequently Asked Question*, les Foires Aux Questions (FAQ) sont ce que l’on pourrait décrire comme des archives contenant les réponses aux questions les plus récurrentes posées par de nombreux utilisateurs. A la différence des forums ou des *Q&A*, « FAQs are created and maintained by experts and the quality of the questions and the answers is good » (Jeon, Croft, and Lee 2006). En d’autres termes, les données des FAQ sont en général de bonne qualité et peuvent être utilisées en vu de constituer un corpus pouvant servir à un système QR. Pour résumer « The idea behind a FAQ is to record the consensus of opinion among a group on some common question and make that answer available, particularly to newcomers who may otherwise ask the same questions again and again. » (Burke et al. 1997). Cette assertion nous permet aussi de comprendre la problématique des FAQ, qui réside dans le fait que, la plupart du temps, les utilisateurs, et en particulier les nouveaux utilisateurs, ne prennent pas le temps de les explorer et préfèrent obtenir la réponse à leurs questions par d’autres moyens (par exemple en contactant directement l’organisme our les personnes compétentes).

## Problématique et objectifs

Dans le contexte des systèmes de question-réponse et des FAQ, notre problématique s’articule autour du *sentence pair matching*, c’est-à-dire l’appariement question utilisateur – question FAQ. En d’autres termes, il s’agit de laisser à l’utilisateur la liberté de poser à un système une question en langage naturel. Ledit système va rechercher dans sa base de données composée de questions posées dans des FAQ la question la plus similaire à la requête de l’utilisateur. Le système lui renverra ensuite la réponse correspondante.

Notre objectif final est d’implémenter un système de question-réponse à destination de l’Etablissement Français du Sang. Ce système devra répondre aux questions des usagers, donneurs potentiels. Nous considérons que les questions seront posées sous formes écrites en langue naturelle dans une interface dédiée. Nous supposons que les utilisateurs du système QR n’auront pas de connaissances préalables ou particulières dans le domaine médical. Ainsi, les réponses du système devront être vulgarisées afin d’être accessibles pour un large public. C’est ce principe de vulgarisation qui fait la force des FAQ puisqu’il s’agit, à la base, de questions posées par des utilisateurs lambdas que des experts ont quelque peu formatées puis auxquelles ils ont répondu.

## Etat de l’art

#### SQR dans le domaine de la santé.

Quand il s’agit de rechercher des informations dans le domaine médical, les informations sur internet doivent être utilisées avec précaution. En effet, l’utilisateur lambda n’est la plupart du temps pas à même de juger de la pertinence et de la fiabilité des informations médicales qu’il peut trouver en ligne. C’est sur cette problématique que la fondation suisse *Health on the Net Foundation[[1]](#footnote-2)* s’est penchée. (Cruchet, Gaudinat, and Boyer 2008) développent un SQR dans ce domaine. Leur hypothèse est qu’une question sur le thème de la santé est caractérisée par son type médical (symptômes, traitement, causes etc) et par le type de réponse attendu (booléenne, causale, spatiale etc). Le corpus utilisé est composé d’une centaine de questions collectées via des FAQ (*Frequently Asked Questions*) de divers forums spécialisés dans la santé. Contrairement aux SQR déjà existants, les questions n’ont pas été analysées suivant un *pattern-matching* mais par un apprentissage automatique supervisé, où des experts ont classifié une partie des questions pour créer le corpus d’entraînement. Aucune comparaison des deux méthodes n’est cependant faite ; nous ne savons donc pas pourquoi celle-ci a été privilégiée. Un problème majeur a également été rencontré : celui de la taille du corpus. Dans le corpus d’entraînement, certaines classes étaient fortement sous-représentées par rapport à d’autres, ce qui a faussé leur classification.

Une approche concernant la détection de questions similaire développée par (Abacha and Dina 2016) repose sur la reconnaissance de *text entailement*, c’est-à-dire de la relation directe entre deux segments de texte dont le sens de l’un est contenu dans le sens de l’autre.

Q1 (question d’un utilisateur) « A partir de quel âge est-ce que je peux donner mon sang ? »

Q2 (FAQ) « Puis-je donner avant mes 18 ans ? »

Q1 ↔ Q2.

Ce modèle de détection de la similarité est selon nous pertinente dans le sens où les questions posées par des utilisateurs dans des FAQ ou des forums sont rarement directes, peuvent être grammaticalement ou orthographiquement incorrectes et peuvent prendre plusieurs tournures différentes (« est-ce que », inversion sujet-verbe etc).

(Rinaldi et al. 2004) ont, quant à eux, adapté un système QR déjà existant, à la base développé pour un domaine restreint, au domaine de la génomique. Leurs recherches mettent en lumière le problème de l’accès et de l’exploitation de données scientifiques, et notamment le fait qu’on ne puisse pas s’attendre à ce que les utilisateurs d’un SQR à domaine restreint soient à l’aise avec la terminologie du domaine en question. Ainsi, leur système ExtrAns est dédié aux systèmes restreints et possédant une terminologie conséquente. ExtrAns a pu être adapté au domaine de la génomique grâce à la collection et à l’analyse de documents traitant cette problématique. Le problème de la terminologie a pu être contourné grâce à l’annotation manuelle d’éléments terminologiques, permettant ainsi une analyse grammaticale correcte. Le domaine du don du sang étant lui aussi restreint, ces recherches vont nous permettre d’aborder (sinon de résoudre) le problème de la terminologie.

(Niu et al. 2003) s’intéressent quant à eux aux différences entre les systèmes question-réponse à domaine ouvert et ceux dédiés au domaine médical. Les systèmes généraux se concentrent majoritairement sur ce que l’on définit dans la langue anglaise comme étant les *wh-questions* (ce qui équivaut plus ou moins en français à la méthode QQOQCCP: Quoi, Qui, Où, Quand, Comment, Combien, Pourquoi). Dans les systèmes médicaux en revanche, l’emphase est mise sur les traitements d’une maladie, les symptômes, les effets secondaires d’un traitement, etc. Ainsi, l’identification du type de question doit se faire différemment. De plus, le pronom relatif *quand* n’aura pas la même signification dans la question vers un système général que vers un système médical.

SG : *quand a eu lieu la bataille d’Austerlitz ? Le 02 décembre 1805*  date précise

SM : *quand est-ce que je pourrais redonner ? dans 2 mois si vous êtes un homme, 3 si vous êtes une femme*  durée relative

Ils évoquent également le problème des sources de données. Pour qu’un système général soit performant, la réponse doit être validée grâce à plusieurs réponses candidates qui sont redondantes dans la base de données. Dans un système médical en revanche, il est très peu probable qu’une réponse se répète à plusieurs endroits dans la base de données. Cette assertion est à relativiser dans le cas de notre système.

#### Les systèmes QR basés sur des FAQ

A l’inverse des systèmes question réponse traditionnels, les SQR basés sur des FAQ ne répondent pas aux questions en partant d’un corps de données hétérogènes. Les systèmes *FAQ-based* vont plutôt tenter de trouver dans leur base de données une question similaire à la requête d’un utilisateur. S’il y en a effectivement une, la réponse correspondante est alors retournée à l’utilisateur. Ces dernières années, de nombreux modèles SQR basés sur des FAQ ont vu le jour. Ces modèles présentent tous des architectures bien distinctes et ont chacun des méthodes de calcul de similarité différentes (nous reviendrons sur ce point plus tard).

(Wang and Ittycheriah 2015) base leur modèle sur l’alignement lexical (*word alignement*). Dès qu’un utilisateur fait une requête, une question candidate est extraite de la base de données. Une matrice de similarité est créée pour chaque paire de mots entre la requête utilisateur et la question candidate (via une mesure de distance cosine). Le meilleur alignement lexical est ensuite calculé afin d’obtenir un vecteur. Ces différents vecteurs nourrissent un réseau neuronal qui classifie et ordonne les questions de la FAQ de la plus à la moins similaire à la requête utilisateur. Bien que leur modèle génère de bons résultats, un alignement au niveau lexical provoque selon nous une trop grosse perte sur le plan sémantique. Nous reviendrons sur cette assertion plus tard.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 | C5 | C6 |
| Q1 | 0.7 | 0.2 | 0 | 0.1 | 0 | 0.2 |
| Q2 | 0.1 | 0.5 | 0 | 0.2 | 0 | 0.3 |
| Q3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q4 | 0 | 0 | 0.2 | 0.1 | 0.6 | 0 |
| Q5 | 0 | 0.2 | 0 | 0.1 | 0.3 | 0.8 |

*Figure 2. Matrice de similarité et alignement textuel comme décrit par Wang et Ittycheriah.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 | C5 | C6 |
| Q1 | X |  |  |  |  |  |
| Q2 |  | X |  |  |  |  |
| Q3 |  |  |  |  |  |  |
| Q4 |  |  |  |  | X |  |
| Q5 |  |  |  |  |  | X |

(Burke et al. 1997) base, quant à eux, leur modèle, dénommé FAQ Finder, non seulement sur la similarité métrique entre les documents, mais aussi sur la base de connaissance WordNet. Leur système procède en plusieurs étapes : d’abord, la question de l’utilisateur est *pré-traitée* : les mots vides sont retirés et la requête est vectorisée. Il est intéressant de noter que, contrairement à la plupart des systèmes de recherche d’information, ils procèdent à une analyse morphologique des mots, et non pas une ‘‘racinisation’’ (*stemming*). La seconde étape consiste à *faire correspondre* ce vecteur aux les autres questions présentes dans la FAQ, elles aussi vectorisées. Un score est attribué à chaque paire (vecteur\_utilisateur, vecteur\_FAQ). Ce score est calculé grâce à la similarité statistique ainsi qu’à la similarité sémantique. La première est calculée à l’aide de la mesure *tf-idf* puis de diverses mesures de similarité (cosine, Dice, Tanimoto et coefficient de Szymkiewicz–Simpson), notions sur lesquelles nous reviendrons plus tard. La similarité sémantique est, quant à elle, calculée grâce à WordNet, une base de connaissance constituée, entre autres, de groupes de synonymes et de relations lexicales, telles que l’hyponymie. La qualité des résultats produits par ce système est due à l’utilisation de deux types de similarité, statistique et sémantique.

## Concepts clés autour des systèmes question réponse

Dans un système question-réponse, la question doit tout d’abord passer par un processus de normalisation. (Jurafsky and Martin 2017) définissent trois étapes indispensables à la normalisation de texte.

* La segmentation (ou tokenisation) des mots. Selon (Clark, Fox, and Lappin 2010), il s’agit de ‘casser’ les séquences de caractères présentes dans un segment de texte afin de localiser les limites d’un mot. Dans cette optique, deux types de langues se distinguent : celles séparés par un espace et celles qui ne le sont pas. Cette approche met en lumière le problème de la ponctuation (par exemple les mots séparés par un tiret, notamment dans les structures interrogatives) et celui des expressions comportant plusieurs mots.
* Normaliser les tokens présents dans le texte, c’est-à-dire le processus de lemmatisation, défini par (Clark, Fox, and Lappin 2010) comme étant la détermination de la racine d’un mot.
* La segmentation des phrases. Ce processus s’appuie le plus souvent sur la ponctuation. La présence d’abréviations tend à poser un problème puisque le point pourrait alors être mal analysé.

Une fois le texte normalisé, il passe par un processus d’analyse qui comprend les étapes suivantes, définies par (Embarek, 2008).

* Le type de question. Il permet de classifier les questions dans des types comme booléen, définitoire, etc. Ainsi, la question « comment se déroule un don ? » sera étiquetée comme question procédurale tandis que la question « pourquoi ne peut-on pas donner en dessous de 50kg ? » se verra attribuer la catégorie de question causale.
* La reconnaissance d’entités nommées. (Omrane, Nazarenko, and Szulman 2009) définissent les entités nommées (EN) comme des éléments qui soulignent les fonctions référentielles de certains segments de texte. Ainsi, les EN agissent en quelque sorte comme des noms propres puisqu’elles font référence à des lieux, des personnes ou encore des dates.
* Le type de réponse attendue. (Mendes and Moriceau 2004) ont dégagé dix types principaux, à savoir (1) vériconditionnelle, (2) catégorie conceptuelle, (3) quantitative, (4) qualitative, (5) procédurale, (6) définition, (7) comparaison et/ou évaluation, (8) causale, (9) but et (10) description. (Li and Roth 2006) ont quant à eux défini une taxonomie comprenant cinq grandes catégories et cinquante classes plus précises (voir annexes).

Le type de réponse attendu dépend évidemment du type de la question et des entités nommées présentes dans la question.

* Le focus de la question. Il permet de savoir sur quel élément on doit fournir des informations dans la réponse. Par exemple, « quel est le poids minimal requis ? » aura comme focus le segment textuel *poids*.

Toutes ces étapes utilisent différents outils tels que des *parsers*, des *pattern-matchers*, des *POS-taggers* ou encore des annotations sémantiques, ces processus pouvant varier en fonction du système développé.

Dans le schéma de traitement général adopté par la majorité des systèmes question-réponse, une fois la question finement normalisée et analysée, le système interroge un corpus qui est composé de bases de données hétérogènes afin d’en extraire les passages textuels contenant potentiellement la réponse à la question. Ces passages vont ensuite être segmentés afin d’analyser les phrases une par une et de les classer. Le classement peut être effectué selon différents critères, dont une liste non-exhaustive est présentée par (Jurafsky and Martin 2017).

* Le nombre d’entités nommées du bon type dans le passage (si la question attend une réponse de type DATE, les passages du corpus contenant le même type d’entité seront extraits comme des réponses potentielles)
* Le nombre de mots-clés de la question contenu dans le passage.
* La plus longue séquence de mots de la question qui apparaît à l’identique dans le passage (Qutilisateur : « **la prochaine collecte de sang a lieu** quand ? » / Pcorpus : « **la prochaine collecte de sang aura lieu** le [date] »)

La réponse classée première en fonction du ou des critères choisis est alors extraite du corpus puis renvoyée à l’utilisateur.

# Méthodologie pour l’appariement de questions dans un système QR basé sur une FAQ

## Hypothèses de recherche

L’une de nos premières hypothèses est que l’orthographe joue un rôle capital dans un système question-réponse, qu’il soit basé ou non sur une FAQ. Ceci est d’autant plus important que le système que nous cherchons à construire est à destination du grand public. Si le segment textuel saisi par l’utilisateur comporte des erreurs de graphie, le module de lemmatisation ne parviendra pas à ramener les mots sous forme de lemmes. Dans ce cas, la requête sera mal interprétée, la question mal modélisée, et la réponse extraite ne correspondra certainement pas à la question posée par l’utilisateur. Pour pallier ce problème, l’utilisation de la distance de Levenshtein est envisagée pour corriger les saisies de l’utilisateur.

Notre deuxième hypothèse est qu’une segmentation ainsi qu’une lemmatisation automatique et des analyses morpho-syntaxiques, en utilisant des librairies logicielles existantes, produira des résultats satisfaisants pour ce qui est de l’étape du *pre-processing* de la question. Nous avons prévu d’utiliser deux modules, NLTK[[2]](#footnote-3) et TreeTagger[[3]](#footnote-4), implémentés en Python. Le premier module, NLTK, permet notamment de segmenter, d’étiqueter, de lemmatiser, d’identifier les entités nommées et de construire des arbres de dépendance. Le deuxième module, TreeTagger, permet d’étiqueter les *parties du discours* et d’obtenir les lemmes.

Notre troisième hypothèse est que, pour effectuer un appariement correct des questions, le calcul d’une seule mesure de similarité est insuffisant. Nous envisageons donc d’utiliser diverses mesures (tant statistiques que sémantiques) afin d’obtenir des performances optimales. Nous procéderons à une évaluation du modèle afin de déterminer quelle(s) mesure(s) sont les plus performantes.

Enfin, notre quatrième hypothèse est qu’une vectorisation à l’échelle lexicale de notre corpus et de la requête utilisateur est insuffisante et provoque une forte perte d’information sur le plan sémantique. Nous proposons donc, en plus d’une telle analyse, une vectorisation à l’échelle phrastique, ce qui nous permettra de limiter la perte sémantique.

Nous proposons donc, à partir d’un corpus de questions posées par des utilisateurs moissonné dans des FAQ, de construire un algorithme qui permet d’obtenir la représentation vectorielle d’une question. En parallèle, il s’agira de représenter le corpus de la même manière, de façon qu’une comparaison puisse être faite. C’est pourquoi nous allons également construire une base de données nous permettant d’exploiter correctement le corpus.

## Domaine

Le domaine de notre SQR sera …. L’objectif principal est de ...

Cet objectif nous tient à cœur car selon nous, l’une des raisons qui pourrait expliquer la baisse des réserves de sang en France est la mésinformation des citoyens. Grâce à ce système de question-réponse, nous espérons simplifier l’accès aux informations concernant le don du sang et ainsi inciter plus de personnes à le faire. Il arrive que certaines personnes se rendent à une collecte dans le cadre d’un don, mais soient refusés du fait qu’elles ne remplissent pas certaines conditions. Ces conditions devraient pouvoir être non seulement connues mais comprises à l’avance.

De plus, le site de l’EFS[[4]](#footnote-5), bien qu’il contient toutes les informations relatives au don du sang, ne bénéficie pas d’une interface ergonomique ; plusieurs rubriques portent par exemple le même nom mais ne sont pas situées au même endroit. Un système QR permettrait de regrouper toutes ces informations et de les rendre accessbiles à partir d’une seule et même interface, simplifiant ainsi l’accès pour les utilisateurs.

Le domaine du don de sang, très restreint, fait partie du domaine plus général de la santé. A notre connaissance, aucun système question-réponse n’ait jusqu’ici été développé dans ce domaine

## Ressources

#### Corpus de questions

En ce qui concerne la construction du corpus de questions, nous allons en premier lieu exploiter les ressources présentes sur le site internet de l’Etablissement Français du Sang et les regrouper sous forme de catégories. En rassemblant les différentes rubriques, nous obtenons les catégories suivantes :

* **Critères principaux** : cette catégorie concerne l’âge, le poids etc.
* **Etat de santé**: cette catégorie concerne les contre-indications au don, les antécédents médicaux etc.
* **Mode de vie** : cette catégorie concerne la sexualité, la consommation d’alcool et/ou de drogue, les voyages à l’étranger etc.
* **Déroulement d’un don** : cette catégorie concerne la durée d’un don, la quantité de sang prélevée etc.

Nous envisageons de construire ce corpus en représentant les données sous forme d’un document XML qui aurait comme balises les différentes catégories citées ci-dessus ainsi que les sous-catégories qu’elles incluent. Cela faciliterait l’identification des questions .

<criteres\_principaux>

<age>puis-je donner avant mes 18ans ?</age>

<poids>puis-je donner si je pèse 45 kg ?</poids>

Les questions présentes dans les FAQ du site de l’EFS sont présentées en langue ‘écrite’ tandis que nous supposons que les questions posées par les utilisateurs seront présentées en langue ‘orale’ . Ainsi, nous doutons que les utilisateurs se servent de la tournure inversion sujet-verbe lorsqu’ils soumettront une question au système alors que c’est sous cette forme qu’elles se trouvent dans les FAQ. De cet fait, il nous faudra nous baser sur les recherches sur le *text entailement* d’(Abacha and Dina 2016).

En exploitant uniquement les questions présentes dans les FAQ du site de l’EFS, nous obtenons un corpus de 34 questions, ce qui est peu. En revanche, en additionnant les différentes questions présentes sur les différents sites cités dans l’annexe 2, nous obtenons un corpus plus conséquent de 179 questions. L’enrichissement de ce corpus a également été possible grâce à des FAQ en langues étrangères, que nous avons traduit en français. La redondance des questions est évidemment un atout, puisque que nous avons ainsi plus de tournures phrastiques portant sur le même sujet. Ainsi, cela augmente les chances que la requête utilisateur soit déjà présente dans la base de données avec très peu, voire pas du tout de variations.

#### Corpus de réponse

Le corpus de réponse sera, quant à lui, construit en exploitant les ressources présentes sur le site de l’EFS, mais également les ressources présentes sur d’autres sites internet (des exemples sont cités en annexe). Ce corpus doit être construit avec vigilance et rigueur car une mauvaise réponse dans ce domaine peut avoir de lourdes conséquences lors de l’utilisation du système. Ainsi, nous nous restreindrons à des sites internet spécialisés dans ce domaine et vérifiés en tant que sources fiables afin d’assurer la véracité des réponses fournies.

## Types de questions

Dans cette partie, nous dressons la liste des types de questions que nous avons observés dans les FAQ de l’EFS.

#### Les questions fermées.

Elles attendent une réponse de type booléen, c’est-à-dire une validation ou au contraire une invalidation. Ces dernières peuvent être de type *nécessairement vrai* (la nécessité) ou bien de type *contingentement vrai* (la possibilité)– ces notions proviennent de la logique aléthique (Alliot et al. 2002). Un énoncé est *nécessairement vrai* s’il est validé ou invalidé en permanence. Par exemple, la proposition A « on ne peut jamais donner son sang si l’on a moins de 18ans » est *nécessairement vrai*; on écrira alors □A. En revanche, un énoncé est *contingentement vrai* s’il n’est validé que dans certaines conditions. Par exemple, la proposition B « je dois emmener des documents spécifiques lors du don ? » est *contingentement vrai* puisque sa validation dépend de s’il s’agit d’un premier don ou non ; on écrira alors ◇B. Ainsi, au niveau sémantique, une réponse à une question fermée fait partie de l’ensemble E {p, p, □p, ◇p, □p, ◇p} (Langlet and Clavel 2015).

Au niveau morpho-syntaxique, en français, les questions fermées peuvent être représentées de différentes façons.

* **L’inversion sujet-verbe**. Comme dans de nombreuses langues, le mode interrogatif de la langue française demande une inversion sujet-verbe. Cependant, cette tournure tend de plus en plus à être synonyme d’un haut niveau de langue et donc à être peu utilisée dans la langue parlée. *Cela pose-t-il un problème ?*
* **« est-ce que »**. Etymologiquement, cette expression vient de l’inversion de « c’est que », mais elle n’a aujourd’hui plus la même signification. Cette tournure est largement utilisée dans les structures interrogatives.

*Est-ce que cela pose un problème ?*

* **La structure conditionnelle**. Une structure conditionnelle peut être utilisée dans le cadre d’une demande d’information. Techniquement, elle n’appelle pas de marque d’interrogation.

*Je me demandais si cela posait un problème.*

* **Une simple marque d’interrogation**. A l’oral, cette structure se manifeste par une intonation montante qui est appliquée à une phrase déclarative. A l’écrit, il s’agit simplement de mettre un point d’interrogation à la fin de la phrase pour la convertir en question.

*Cela pose un problème ?*

Il faut cependant tenir compte du fait que la réponse à une question fermée ne se limite pas à « oui » ou « non ». Dans la majorité des cas, des informations supplémentaires permettant d’expliquer la polarité de la question seront nécessaires. En effet, le principe du système QR que nous cherchons à construire est de pouvoir apporter une réponse **complète** à l’utilisateur ; il ne doit donc pas se contenter de réponse positive ou négative, mais d’expliquer pourquoi il a répondu de cette manière.

#### Les questions ouvertes

Par opposition aux questions fermées, les questions ouvertes sont celles qui n’attendent pas de réponse booléenne ; elles sont au contraire totalement libres. La réponse à une question ouverte est liée aux mots-clés et/ou aux entités nommées présentes dans la question.

Exemple :

Qutilisateur: « Combien faut-il peser pour donner son sang ? »

Dans l’exemple de question ouverte ci-dessus, le système devra reconnaître l’entité nommée *combien* et l’associer au verbe *peser*. Ainsi, les passages de réponses à extraire contenant une réponse potentielle à cette question doivent contenir des informations sur un poids. Cependant, ce qui s’applique aux questions fermées s’applique aussi aux questions ouvertes. Si l’on reprend l’exemple ci-dessus, le système ne peut pas se contenter de retourner une réponse de type poids, par ex. 50kg; il doit aussi retourner un passage, c’est-à-dire un texte explicatif qui puisse justifier la réponse retournée.

# Théories linguistiques

## La théorie de la sémantique (?)

“The most characteristic feature of language is its ability to make available an infinity of sentences”

(Katz and Fodor 1963)

Selon Jerold Katz et Jerry Fodor, la caractéristique sémantique de la langue n’atteint pas le niveau de théorisation de la caractéristique grammaticale de celle-ci. Sur une échelle grammaticale, les catégories morpho-syntaxiques sont figées par exemple ; il n’en existe pas une infinité. Sur une échelle sémantique en revanche, on peut considérer que le sens ainsi que la manière dont il est véhiculé n’a pas de limites. (Katz and Fodor 1963) expliquent que “the striking fact about the use of language is the absence of repetition: almost every sentence uttered is uttered for the first time”. En effet, il faut ici considérer le langage par sa fonction communicative : un locuteur utilise la langue pour faire passer un message, c’est-à-dire pour s’exprimer et communiquer. C’est d’ailleurs les locuteurs d’une langue qui définissent la sémantique de cette dernière selon (Harris 1968) : “to say that a property P of expressions is semantic is just to say that competent speakers decide whether an expression has P or not based on the meaning of P”. Cette assertion remet cependant en perspective la nécessité syntaxique d’une langue, dont nous reparlerons plus tard.

(Katz and Fodor 1963) énoncent également le principe de compositionnalité sémantique de la langue : “since the set of sentences is infinite and each sentence is a different concatenation of morphemes, the fact that a speaker can understand any sentence must mean that the way he understands sentences which he has never previously encountered is compositional”. Ce principe a été mis en lumière par le philosophe allemand Friedrich Ludwig Gottlob Frege, qui l’énonce dans (Frege 1953) de la manière suivante : “never ask for the meaning of a word in isolation, but only in the context of a proposition”. En d’autres termes, le sens d’un mot doit être considéré uniquement en contexte ; c’est également l’un des principes fondamentaux de la théorie distributionnaliste, que nous abordons plus bas.

## Les limites de la syntaxe

Pour beaucoup de linguistes, la langue naturelle se divise en deux grandes parties : la syntaxe et la sémantique. On peut définir la syntaxe comme étant l’ensemble de règles qui régissent une langue. Pour Chomsky notamment, syntaxe et sémantique sont des concepts opposés, bien qu’il revienne sur cette assertion dans (Chomsky 2002) : “there is no aspect of linguistic study more subject to confusion and more in need of clear and careful formulation than that which deals with the points of connection between syntax and semantics”. Pour (Katz and Fodor 1963), l’utilisation de grammaire à des fins descriptives en ce qui concerne la langue naturelle est limitée.

First, a grammar cannot account for the fact that some sentences which differ ONLY morphemically are interpreted as different in meaning (e.g. *The tiger bit me* and *The mouse bit me*) while other sentences which differ only morphemically are interpreted as identical in meaning (e.g. The oculist examined me and The eye doctor examined me). Second, a grammar cannot account for the fact that some sentences of radically different syntactic structure are synonymous (e.g. *Two chairs are in the room* and *There are at least two things in. the room and each is a chair*) while other syntactically different sentences are not. In each case, the interpretation of the sentences is determined in part by the meanings of their morphemes and by semantic relations among the morphemes.

Ce passage est particulièrement intéressant puisque nous pouvons directement le mettre en lien avec notre problématique. Nous avons, dans notre corpus, beaucoup de structures qui diffèrent uniquement à l’échelle d’un seul morphème et dont le sens est radicalement différent (par exemple, « je suis enceinte » et « je suis diabétique »). A l’inverse des structures peuvent être parfaitement synonymes (« puis-je donner si je suis homosexuel » et « puis-je donner si je suis gay »). Nous avons aussi des phrases similaires sur le plan sémantique dont la structure syntaxique est totalement différente (« Je viens de me faire tatouer. Puis-je faire un don » et « Puis-je donner après avoir fait un tatouage ou un piercing »). Ainsi, nous n’accorderons que peu d’importance à la structure syntaxique des phrases. L’hypothèse d’un *pattern-matching* syntaxique est exclu, puisqu’en reprenant l’un des exemples cités ci-dessus, nous obtenons le même schéma syntaxique, bien que le sens soit discordant.

je PRO:PER je PRO:PER

suis VER:pres suis VER:pres

enceinte ADJ diabétique ADJ

## La paraphrase et la Théorie Sens-Texte

#### Fondements de la Théorie Sens-Texte

“Sans paraphrase, la linguistique serait appauvrie au point où l’on pourrait se demander s’il vaut la peine d’être linguiste !”

(Mel’čuk 1992)

De par cette assertion, Mel'čuk résume les fondements de la Théorie Sens-Texte (TST). La fondation de la TST remonte aux années soixante, sous sa propre impulsion ainsi que celle d’Aleksandr Žolkovskij et de Juri Apresjan. La Théorie Sens-Texte établit que pour une langue 𝓛, il existe une (quasi) infinité d’énoncés 𝓔 de véhiculer un même sens 𝓢, ce que théorisait déjà (Frege 1892) : ‘‘the same sense has different expressions in different languages or even in the same language’’. Ainsi, on peut résumer la TST ainsi :

Il convient alors d’aborder un énoncé sous deux formes : sémantique (c’est-à-dire ayant un contenu informationnel) et textuelle (c’est-à-dire la forme linguistique sous laquelle il apparaît). Ainsi, Mel'čuk définit que le sens d’une phrase doit être considéré sous trois aspects :

* le sens situationnel, qui « représente l’état de chose dont il s’agit » ; c’est ce que l’on définit comme l’aspect sémantique de la phrase.
* le sens communicatif, qui « représente l’organisation du message par le locuteur ». Cet aspect peut être mis en parallèle de la notion de thème topical introduite par (Halliday 2004) dans sa théorie de linguistique systémique. En d’autres termes, il s’agit de ce que le locuteur a décidé de mettre en avant dans son énoncé, que nous appellerons l’aspect organisationnel de la phrase.
* le sens rhétorique, qui « représente les effets expressifs ou artistiques visés par le locuteur ». Cet aspect fait par exemple appel aux niveaux de langues (familier, courant, soutenu), aux figures de style telles que l’ironie

Nous laisserons de côté le dernier sens puisque nous nous intéressons au sens véhiculé, et non à la manière dont il est véhiculé. Ainsi, pour un énoncé 𝓔 donné, le sens situationnel restera le même ; le sens communicatif sera l’élément variant.

Mel'čuk met également en avant un aspect assez contre-intuitif de la paraphrase appelé neutralisation sémantique contextuelle, qui établit que des paraphrases peuvent être construites à partir d’unités linguistiques sémantiquement non synonymiques, voire opposées. Par exemple, « tout » et « chaque » peuvent sembler sémantiquement opposés. Pourtant, « tout homme est mortel » et « chaque homme est mortel » sont des énoncés paraphrastiques puisqu’ils possèdent tous deux le même sens situationnel.

Il est cependant important de noter que Mel'čuk ne prend en compte que les paraphrases langagières (que lui-même appelle paraphrases linguistiques), c’est-à-dire les paraphrases dont la reconnaissance est rendue possible uniquement grâce à des marqueurs linguistiques. Ces paraphrases langagières sont à opposer aux paraphrases cognitives, c’est-à-dire les paraphrases pour lesquelles des connaissances extralinguistiques sont nécessaires afin d’être reconnues. Il admet cependant l’existence de deux types de paraphrasage au sein des paraphrases linguistiques : les paraphrases sémantiques et les paraphrases syntaxiques. Les paraphrases sémantiques agissent à l’échelle lexicale (essentiellement à un niveau synonymique) tandis que les paraphrases syntaxiques agissent, comme leur nom l’indique, sur la syntaxe. Nous considérons cependant les paraphrases sémantiques comme étant à mi-chemin entre les paraphrases langagières et les paraphrases cognitives dans le sens où la détection de synonyme fait plus ou moins appel à la cognition. Par exemple, « il a la capacité d’ubiquité » et « il peut être partout à la fois » sont, selon la TST, des paraphrases sémantiques, encore faut-il détecter la relation entre « ubiquité » et « pouvoir être partout à la fois ».

#### Les composantes du Modèle Sens-Texte

Ainsi, dans son Modèle Sens-Texte, Mel'čuk définit six composantes principales : la sémantique, la syntaxe de surface, la syntaxe profonde, la morphologie de surface, la morphologie profonde et enfin la phonologie. Dans le cadre de nos recherches, nous ne nous intéresserons qu’à la composant sémantique.

« La théorie Sens-Texte pose comme tâche primordiale une description formelle et exhaustive de la synonymie de phrases, c'est-à-dire de la richesse paraphrastique. Cette description présuppose, à son tour, la description de l'invariant sémantique des paraphrases observables […]. »

(Mel’čuk 1992)

De ce fait, l’un des buts premiers de la TST consiste à formaliser le sens 𝓢 d’un ensemble 𝓔 d’énoncés, c’est-à-dire son invariant sémantique. Il est tout à fait possible de faire une vague approximation de la cardinalité de cet ensemble. Prenons par exemple la phrase suivante : « puis-je donner mon sang si je suis anémié ? ». En procédant à une segmentation en tranches sémantiques, nous obtenons le schéma suivant :

(Puis-je)A | (donner)B | (mon sang si)C | (je suis anémié)D ?

Dans le tableau ci-dessous, chaque colonne contient des équivalences (non-exhaustive) plus ou moins synonymiques des éléments composants les différentes tranches sémantiques A, B, C et D.

|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **B** | **C** | **D** |
| puis-je  (est-ce que) je peux | donner  faire don de | mon sang si |  |

Ainsi, nous obtenons les valeurs suivantes : Nous pouvons de ce fait calculer approximativement le nombre total de variations possibles en multipliant le nombre de variations possibles par tranche sémantique :

Pour un même sens 𝓢, nous obtenons un ensemble d’énoncés 𝓔 de cardinalité 162, cette cardinalité étant ici minimale puisque la liste des variations possibles n’est pas exhaustive. Mel’čuk établit ainsi, de manière exhaustive, que la construction paraphrastique s’effectue grâce à des règles syntaxiques ainsi qu’à des règles lexicales qu’il détaille au nombre respectif de 29 et 54.

## La paraphrase en TAL

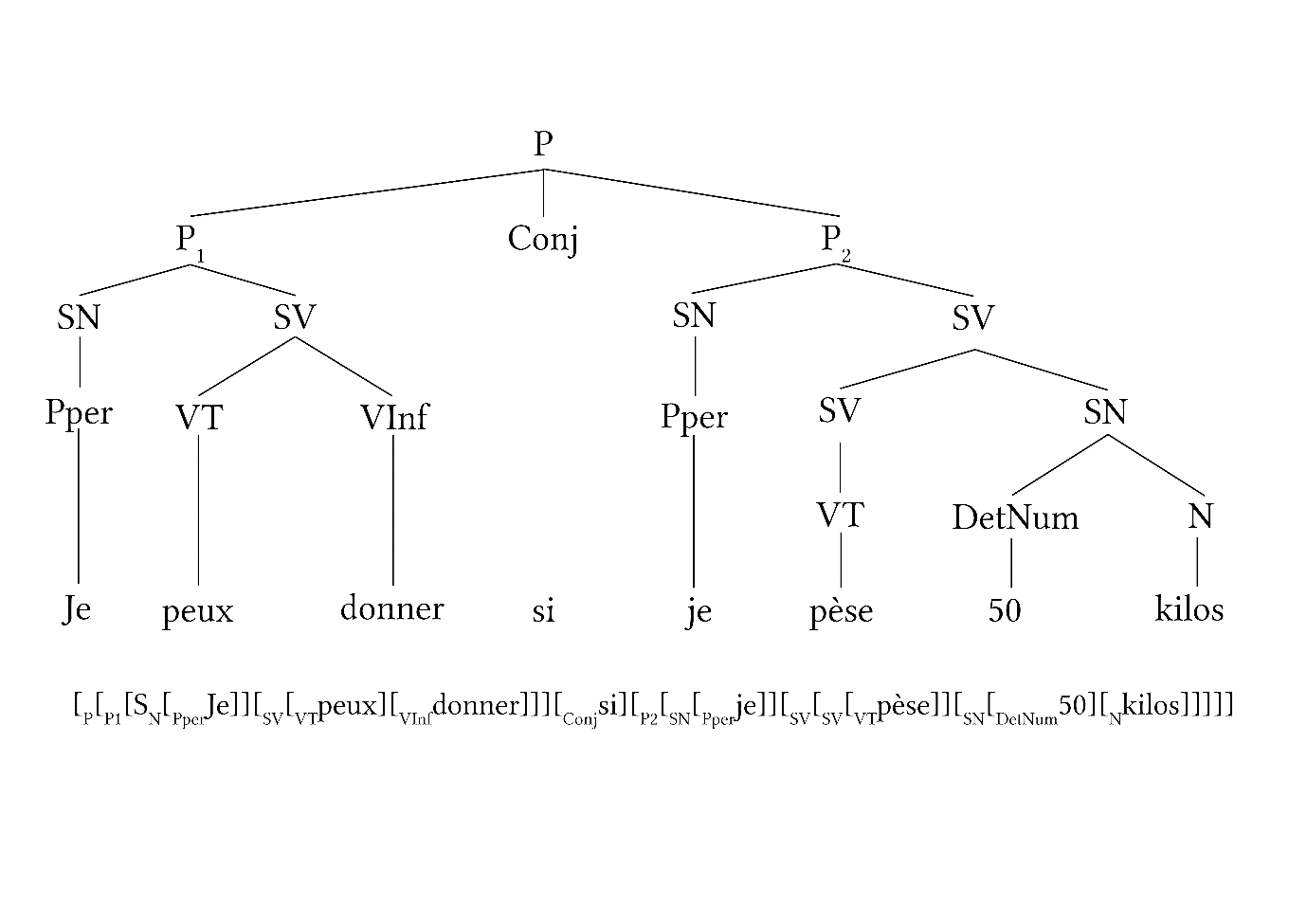
Katz and Fodor abordent dans The Structure of Semantic Theory le problème de la compréhension sémantique par les machines, et plus particulièrement le problème de la compréhension paraphrastique. En effet, pour qu’une machine “comprenne” qu’un segment textuel est la paraphrase d’un autre segment textuel, il faut que ces deux segments comportent “exactly the same syntactic structure and the corresponding words are either identical or synonymous”. C’est pourquoi utiliser les méthodes de tf-idf ou bien la distance de Jaccard ne nous semble pas pertinent, bien que nous justifiions leur utilisation par le fait que dans notre cas, il est difficile de construire des paraphrases du segment textuel “puis-je donner si je pèse moins de 50 kilos ?” sans répéter les éléments “peser” et “kilos”.

## Le distributionnalisme

La théorie générale du distributionnalisme, établit par Leonard Bloomfield et Zellig Harris, consiste à segmenter une phrase en différents éléments et à analyser ces derniers, en examinant notamment comment les éléments “gravitent” et s’organisent entre eux. (Bloomfield 1933) nomme ces éléments “constituents” (constituants). Selon lui, une phrase est une forme complexe composée de plusieurs “ultimate constituents” (composants fondamentaux). Cependant, il statue que “*we could not understand the forms of a language if we merely reduced all the complex forms to the ultimate constituents*”. Illustrons cette assertion par un exemple en langue allemande : Ich glaube nicht, dass der Hund, den ich sehe, ein Deutscher Schäferhund ist. Si l’on décompose la phrase uniquement en constituants fondamentaux, on comprendra la chose suivante : je ne crois pas, que le chien, lui je vois, un berger allemand est, ce qui n’a aucun sens. On remarque de ce fait, comme l’expriment justement (Descles and Biskri 1995), qu’“un énoncé peut être considéré comme une suite de mots ordonnés selon des règles syntagmatiques qui ne reflètent pas forcément l’ordre dans lequel les mots s’appliquent les uns aux autres pour construire l’interprétation sémantique fonctionnelle”.

Ainsi, Bloomfield introduit la notion de “immediat constituents” (constituants immédiats), c’est-à-dire que l’on applique à chaque constituant fondamental formant un constituant immédiat une analyse combinatoire. La théorie de la sémantique combinatoire est abordée en détail plus loin dans nos recherches. Elle établit que l’on peut additionner le sens sémantique individuel des constituants afin d’obtenir le sens total de l’énoncé. Ce découpage en constituants immédiats fait écho à la notion de *tranches sémantiques* introduite par Mel’čuk.

De ce fait, l’analyse dite en constituants immédiats permet de mettre en avant les différentes relations qu’entretiennent les syntagmes à l’intérieur d’une même phrase. Cependant, sur le plan sémantique, cette analyse devient problématique. Considérons la représentation en constituants immédiats de l’énoncé A suivant : *je peux donner si je pèse 50 kilos*.



On observe ainsi les dépendances syntaxiques existantes entre les différents éléments qui composent l’énoncé. Cependant, celles-ci nous en apprennent peu sur les dépendances sémantiques. En effet, dans un arbre syntaxique, on peut intervertir les éléments de même nature (si l’on se situe sur l’axe paradigmatique). Ainsi, selon l’arbre de l’énoncé A (figure XXX), nous pourrions tout à fait construire l’énoncé B \**je pèse 45 kilos si je peux donner* ou l’énoncé C \**je pèse donner si je peux 45 kilos*. Ces énoncés sont syntaxiquement corrects, mais sémantiquement erronés. Cela s’explique par le fait que, bien que *pouvoir* et *peser* appartiennent à la même catégorie grammaticale (verbe), ils n’apparaissent pas dans le même contexte : *pouvoir* apparaît suivi d’un autre verbe à l’infinitif tandis que *peser* apparaît ici suivi d’un numéral (bien qu’il puisse être suivi d’autre chose). C’est exactement ce que théorise (Harris 1968) : “Predicates also have their interpretation conditioned by the nature of their argument”. Son exemple s’appuie sur les différents sens du verbe *cut* (couper) : *to cut one’s finger* (se couper un doigt) signifie « faire un incision » tandis que *to cut prices* (« faire des coupes budgétaires [[5]](#footnote-6)») signifie « réduire ».

# Implémentation du système

## Présentation des librairies utilisées

Avant de pouvoir modéliser les questions afin qu’elles soient traitables par l’ordinateur, nous devons les normaliser, c’est-à-dire effectuer une série de traitements afin de les représenter sous une nouvelle forme adaptée à la recherche de similarités ....

#### Présentation de NLTK et TreeTagger

La bibliothèque NLTK (*Natural Language Toolkit*) est une bibliothèque Python développé en 2005 par Steven Bird et Edward Loper qui permet notamment de segmenter, d’identifier les entités nommées, de construire des arbres de dépendance, d’étiqueter mais intègre aussi une fonction d’apprentissage automatique de classification de textes, ce qui peut être intéressant dans le cadre de notre système.

TreeTagger est un outil développé en 1995 par Helmut Schmid. Cet outil permet l’étiquetage et la lemmatisation de segments textuels. L’atout de cet analyseur réside dans sa capacité à étiqueter et à lemmatiser de nombreuses langues telles que l’allemand, l’anglais, le français, l’italien, le néerlandais, l’espagnol et beaucoup d’autres.

#### Segmentation.

Toute normalisation d’un segment textuel commence par sa segmentation, c’est-à-dire sa division en phrases et/ou en mots en fonction de l’*input (l’entrée ?)* de l’utilisateur. Dans le cas d’une segmentation en phrases, il n’est pas suffisant de considérer la ponctuation (le point, les points de suspension etc) comme marqueur de clôture de phrase, de même qu’il ne siffut pas de considérer la capitalisation d’une lettre comme marqueur de début de phrase.

Q1utilisateur « J’ai des allergies (pollen, graminées, arachides…) et je voulais savoir si c’était un problème dans le cadre d’un don. »

Q2utilisateur « Je suis atteinte de la maladie de Raynaud. Cela pose-t-il un problème ? »

Dans la question Q1, les points de suspension ne marquent pas la fin d’une phrase mais une non-exhaustivité concernant l’énumération qui précède. Dans la question Q2, la majuscule ne correspond pas au commencement d’une nouvelle phrase mais au nom d’une maladie. Il est également à noter que dans le cas de la question Q1, aucune marque d’interrogation (« ? ») n’est visible. Dès lors qu’une demande d’information ou de confirmation d’un fait est formulée, on a affaire à une question.

Lorsque l’on soumet ces deux questions à TreeTagger, la ponctuation et la présence de majuscule ne pose aucun problème. En revanche, les tirets présents dans la question 2 ne sont pas segmentés, ce qui nous donne *-t-il ???.* Lorsqu’on soumet la structure *est-ce que* au système, la même chose se produit. Ce défaut de segmentation n’a cependant pas une grande importance, d’autant plus que TreeTagger est tout de même capable de reconnaître les mots composés comme *porte-manteau*.

#### Lemmatisation.

De même que pour le processus de segmentation, le processus de lemmatisation sera effectué de manière automatique grâce aux modules cités précédemment (NLTK, Stanford CoreNLP, TreeTagger). Nous devons cependant faire face à un problème de taille dans le domaine des systèmes QR : celui de l’orthographe des segments textuels saisis par les utilisateurs. Il serait en effet souhaitable de prendre en compte les erreurs pouvant être commises dans l’*input* d’un utilisateur afin de ne pas omettre un élément important.

Qutilisateur: « je suis atteins**1** d’anémi**2**. C’est grave pour donné**3** son sang ? »

Les éléments 1 et 3 ne poseront pas de problème car un lemmatiseur est capable de traiter des entrées comportant des erreurs de conjugaison. Ainsi, si l’on soumet cette phrase à TreeTagger, on obtient la sortie suivante :

je PRO:PER je

suis VER:pres suivre|être

atteins VER:pres atteindre

d’anémi VER:pper <unknown>

. SENT .

C’est NOM <unknown>

grave ADJ grave

pour KON pour

donné VER:pper donner

son DET:POS son

sang NOM sang

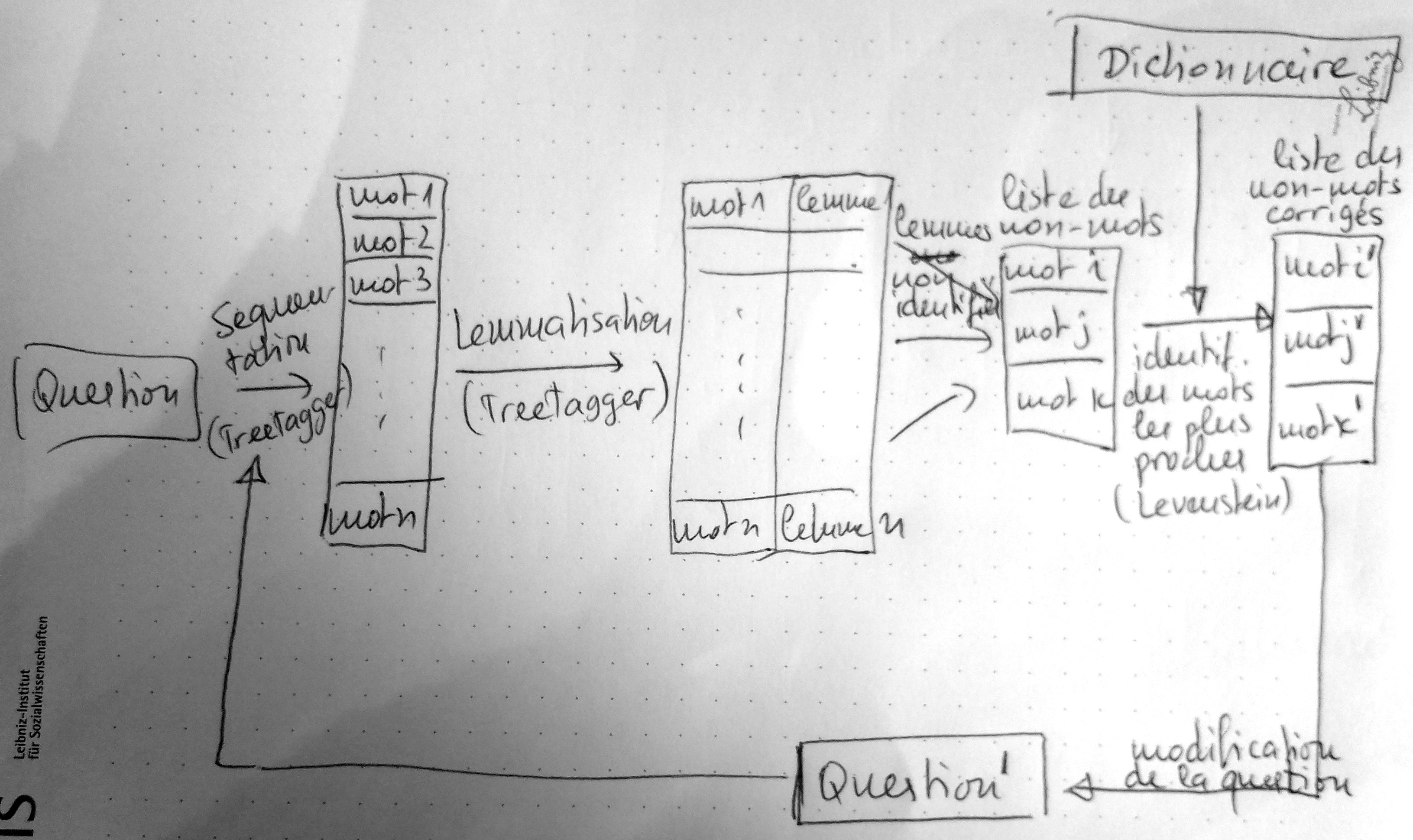
? SENT ?

On se rend alors compte que l’élément 2 est problématique car il présente une erreur d’orthographe qui n’aboutit cependant à aucun autre mot comme c’est le cas pour les éléments 1 et 3. Le lemmatiseur ne parvient alors pas à rétablir sa forme d’origine. (Jurafsky and Martin 2017) définissent deux types d’erreurs :

* La correction de non-mots : il s’agit de la détection et de la correction des erreurs aboutissant à des non-mots (l’omission d’une lettre par exemple, comme l’élément 2 dans l’exemple ci-dessus).
* La correction de mots : il s’agit de la détection et de la correction des erreurs aboutissant néanmoins à des mots existants (comme les éléments 1 et 3 dans l’exemple ci-dessus, qui sont des erreurs cognitives)

Pour corriger les erreurs, nous recherchons tous les mots dans un dictionnaire qui est une ressource externe. Si un mot n’est pas présent dans le dictionnaire, il s’agit de non-mot ; nous utiliserons alors la distance de Levenshtein afin de trouver le mot le plus proche qui appartient au dictionnaire.

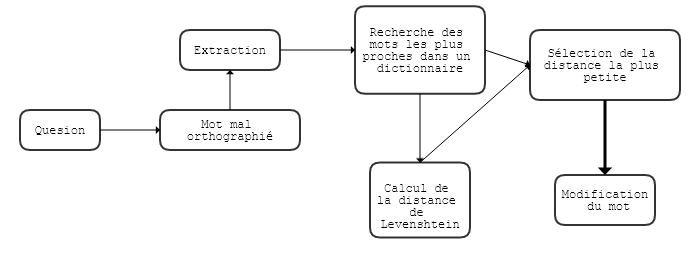
La distance de Levenstein (Citer ici article) se base sur le nombre minimal d’opérations élémentaires de modification de caractères (substitution, addition, suppression) nécessaires à l’obtention de l’un des mots à partir de l’autre. Elle permet, à partir d’une séquence de caractères (ou d’un non-mot), d’identifier dans un dictionnaire les mots les plus « proches », c-a-d les mots qui peuvent être obtenus à partir de l’entrée par un nombre minimal de modifications élémentaires. Nous identifions ainsi le mot qui a la distance de Levenshtein le plus petite avec le non-mot. Dans le cas où il y a plusieurs mots qui ont été retrouvés, ... Cela permettrait de fournir une meilleure entrée pour le lemmatiseur, qui n’aurait pas pu identifier le non-mot à partir de l'entrée initiale. Le schéma général de ce traitement est présenté sur la figure 2.



Par exemple, les mots se rapprochant le plus de l’élément 2 « anémi » pourraient être *anémone*, *anémie*, *ainé*, *année*. Nous obtenons alors les distances suivantes :

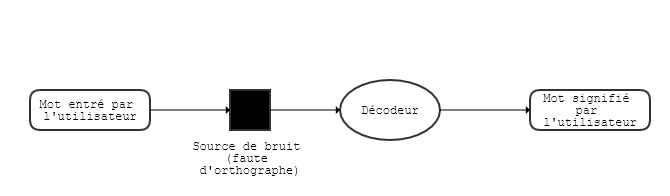
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mot erroné | Mot candidat | Addition | Suppréssion | Modification | Distance |
| anémi | anémone | -one | -i | Substitution et addition | 3 |
| anémi | anémie | -e | \_ | addition | 1 |
| anémi | ainé | -i | -mi | addition et suppression | 3 |
| anémi | année | -n, -e | -mi | substitution, addition, suppression | 3 |

Dans cet exemple, il est clair que le mot signifié par l’utilisateur est alors le mot *anémie*.



*Figure 2. Schéma de modification d’un mot* *mésorthographié grâce à la distance de Levenshtein*

L’autre solution serait d’utiliser l’inférence bayésienne, une méthode permettant de calculer la probabilité d’un évènement, et plus précisément le *noisy channel* *model* développé par(Shannon 1948) car celui-ci peut être appliqué à la correction orthographique. La figure 3 présente le schéma de se traitement.



*Figure 3. Schéma de correction orthographique avec application du modèle de Shannon*

Cependant, le contexte dans lequel cette correction orthographique s’applique est selon nous trop restreint pour que l’utilisation de ce modèle soit pertinente.

Sous-section : Etiquetage morpho-syntaxique

Une fois les segments textuels lemmatisés, nous pourrons procéder à leur étiquetage morpho-syntaxique. Ce dernier est envisagé via différents modules : la bibliothèque NLTK et le logiciel TreeTagger. L’étiquetage morpho-syntaxique nous permettra d’éliminer les éléments dits ‘‘vides’’ (qui servent uniquement d’outils grammaticaux) des éléments dits ‘‘pleins’’ (chargés d’une fonction sémantique) selon la théorie de (Tesnière 1959). C’est sur ces derniers que nous nous appuieront afin d’extraire les passages contenant la réponse potentielle à une question à partir du corpus textuel.

Ainsi, une fois que les différents processus cités précédemment auront été appliqués à la question de l’utilisateur, nous pourrons passer aux étapes suivantes, c’est-à-dire à la vectorisation, puis au calcul de similarité.

## Vectorisation lexicale et phrastique

D’un point de vue mathématique, les vecteurs sont des objets caractérisés par un sens, une direction et une norme (c’est-à-dire une longueur). Ces objets appartiennent à un espace vectoriel, c’est-à-dire à un espace répondant à une loi de composition interne (de type abélien) et à une loi de composition externe. Nous ne nous attarderons pas sur la définition mathématique formelle car ce n’est pas le sujet de nos recherches. D’un point de vue plus macroscopique, un vecteur est un élément servant de support à la transmission d’information. Cette interprétation du sens du mot vecteur nous permet ainsi de faire le lien entre ce dernier et une lexie. En effet, comme nous l’avons étudié plus tôt, c’est la combinaison d’éléments lexicaux qui rend la communication entre deux locuteurs possible. Ainsi, tout comme les vecteurs mathématiques sont projetés dans un espace vectoriel, les éléments appartenant à une langue sont eux même projetés dans un espace ‘‘langagier’’ (réunissant linguistique, sémantique, cognition etc). De ce fait, il ne semble pas contre-intuitif de dire qu’il est possible de vectoriser des unités lexicales telles que des mots, voire des phrases.

### La représentation en *bag-of-words (BOW)*

Très utile dans le domaine de *information retrival*, la représentation en sac-de-mot consiste à représenter un document textuel par un ensemble de termes, en prenant en compte la ‘‘fréquence d’apparition’’ des termes qui le composent. Par exemple, pour effectuer une représentation BOW des phrases P1 « je peux donner si je pèse 50 kilos ? », P2 « je suis apte au don si je suis diabétique ? », P3 « est-ce que je peux donner si je suis enceinte ? » et P4 « combien de temps dure un don ? », il nous faudrait tout d’abord construire ‘‘l’espace lexical’’ (c’est-à-dire le vocabulaire) dans lequel ces phrases sont projetées ; nous obtenons un ensemble de cardinalité 20 : { je, peux, donner, si, pèse, 50, kilos, suis, apte, au, don, diabétique, est-ce, que, enceinte, combien, de, temps, dure, un }. Il s’agit ensuite de observer la fréquence d’apparition de chaque élément de l’ensemble obtenu dans les phrases et contruire le vecteur qui correspond à chaque phrase : P1 (2 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0), P2 (2 0 0 1 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0), P3 (2 1 1 1 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0) et P4 (0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 1 1). Nous pouvons représenter visuellement ces phrases comme le montre la figure XXX.

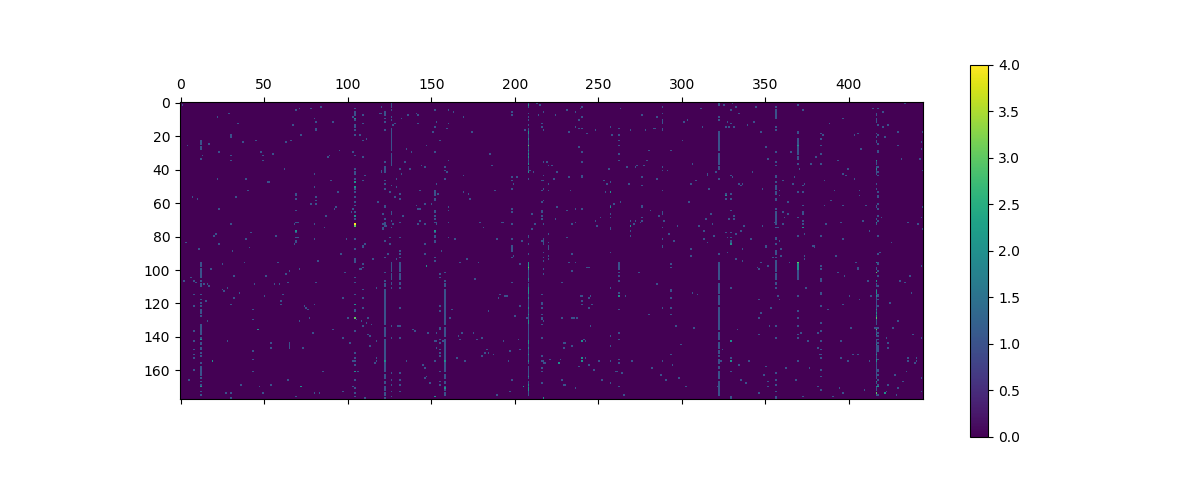


Dans le figure XXX, les éléments apparaissant zéro fois sont représentés en violet, ceux apparaissant une fois en bleu et ceux apparaissant deux fois en jaune. L’idée sous-jacente est que, si deux documents contiennent les mêmes termes, il y a de fortes chances qu’ils soient similaires sur le plan sémantique : Cette affirmation reste approximative, parce que l’ordre d’apparition des éléments dans les textes a en général un effet significatif sur le sens exprimé, et cet ordre n’est pas pris en compte dans la représentation BOW.

C’est sur cette co-occurrence des termes dans des documents que se base la pondération dite tf-idf (*Term Frequency – Inverse Document Frequency*).

Le tf représente la fréquence relative d’un mot ; cette notion a été introduite par Hans Peter Luhn à la fin des années 1950. Son hypothèse était que « the frequency of word occurrence in an article furnishes a useful measurement of word significance » (Luhn 1958). Il est également à noter que, dans sa théorie, la forme des mots, la syntaxe et la sémantique ne sont pas prises en compte ; on se contente alors de faire un ‘‘inventaire’’ (en d’autres termes, une représentation en sac de mot) des lemmes pleins (pour reprendre la dénomination de (Tesnière 1959)) présents. Ainsi, le calcul du tf s’effectue de la manière suivante :, où représente le nombre de fois où un terme apparaît et le nombre total d’éléments présents.

Bien sûr, cette représentation en BOW et le calcul du tf pour chaque terme n’introduit pas de différences significatives à l’échelle des quatre phrases mentionnées précédemment, puisque les mots revenant le plus souvent sont « je » et « suis ». La figure XXX presente la visualisation obtenue lorsqu’on applique la mesure tf à notre corpus composé de 178 questions pour un total de 481 mots différents. L’algorithme implémenté (dont le script se trouve en annexe) nous permet d’obtenir le graphique sur la figure XXX.

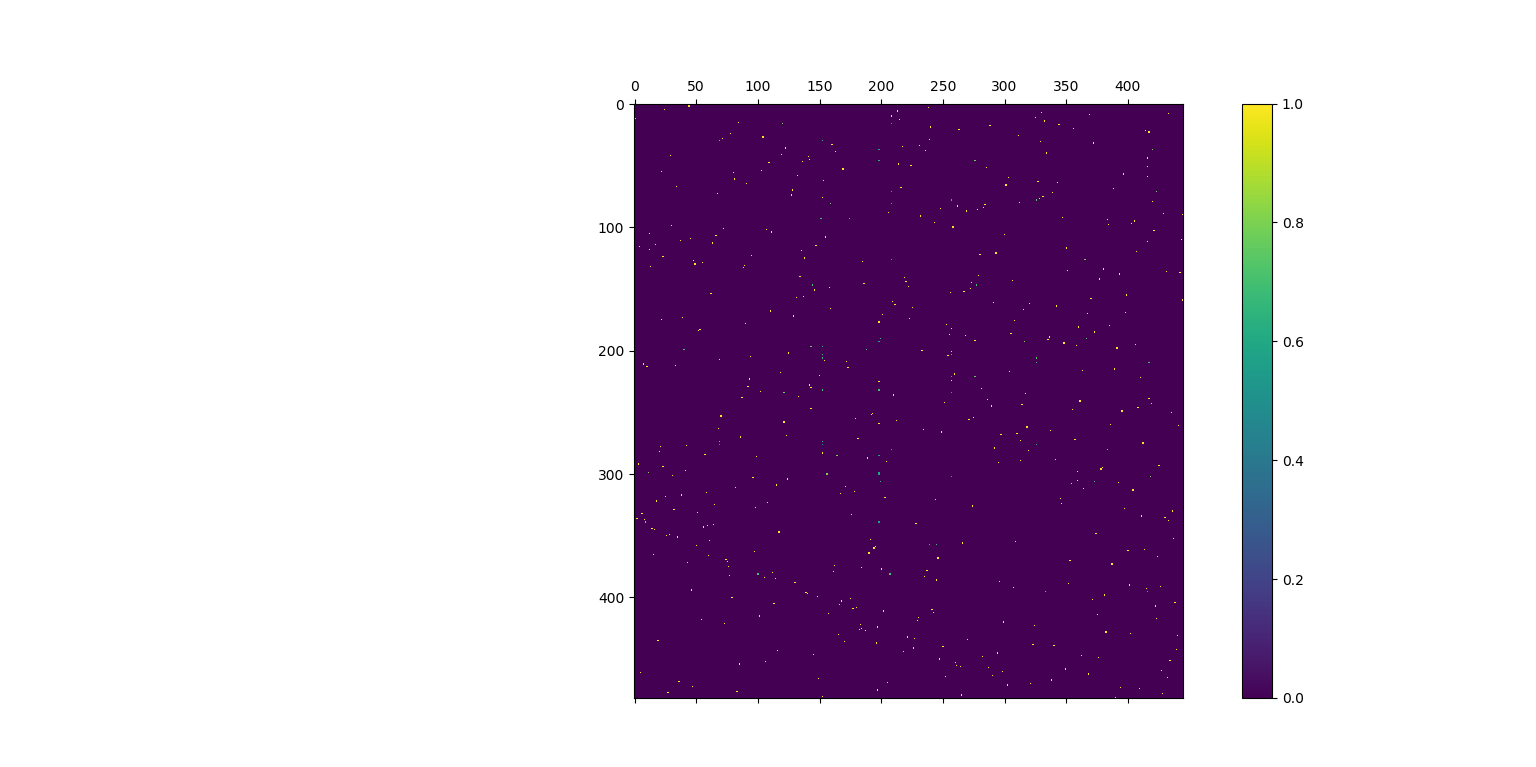


Les lignes bleues verticales que l’on observe sont les mots dont la fréquence d’apparition est relativement élevée .

Nous pouvons émettre l’hypothèse que plus un mot apparaît dans des documents différents (ici des questions), moins il est significatif pour distinguer entre ces documents. Ainsi, il conviendrait de ‘‘pénaliser’’ les mots dont la fréquence d’apparition est trop élevée dans l’ensemble des documents ; c’est ce que l’on appelle l’*Inverse Document Frequency* (idf). Cette notion introduite par Karen Spärck Jones au début des années 1970, représente le poids attribué à un terme pour exprimer son pouvoir discriminatoire. C’est-à-dire que moins le terme apparaît dans les différents documents, plus son poids est élevé ou, pour reprendre les mots de Jones, « very frequently occurring terms [being] responsible for noise in retrieval, one possible course is simply to remove them from requests » (Spärck Jones 1972).

On calcule l’idf de la manière suivante : , avec le nombre total de corpus et le nombre de documents dans lesquels un terme apparaît.

Ainsi, en combinant ces deux notions, on obtient la mesure de tf-idf, qui permet de pondérer la valeur d’un terme dans un document qui appartient à une collection de documents : . L’inconvénient de cette méthode réside dans la nécessité de travailler sur un grand corpus afin d’obtenir des valeurs pertinentes. Ainsi, nous obtenons le graphique suivi :



### 

Cette représentation de l’idf à l’échelle de notre vocabulaire nous montre clairement que, finalement, peu de mots sont significatifs au sein de celui-ci.

Cependant, même sans ce calcul, nous pouvons nous rendre compte de ce phénomène en regardant notre corpus de questions : dans 94 questions par exemple, on retrouve la formule « puis-je faire un don » ou « puis-je donner ». Il est également important de noter que, dans le cadre de nos recherches, l’idf est une notion très importante dans la mesure où il s’agit d’un domaine spécialisé et restreint et elle permet de corriger les poids elevés des termes qui sont répétés uniquement parce qu’il font partie du vocabulaire spécifique du don de sang. Si nous devions enlever de notre corpus les termes, voire les structures, qui apparaissent de manière récurrente, il ne nous resterait que très peu d’éléments textuels à analyser : nous aurions par exemple « piercing » ou « diabétique ». Ainsi, il s’agirait seulement de regarder lequel de ces mots-clés apparaît dans la question posée par l’utilisateur pour l’associer sémantiquement. En même temps, il nous faut prendre en compte la notion de synonymie, ou même d’hyper- et d’hyponymie (par exemple, « piercing » et « tatouage » entre tous deux dans la catégorie « modification corporelles ») afin de faire comprendre au système que, bien que ces mots soient différents, leur sens est lié.

## Les différentes mesures de similarité

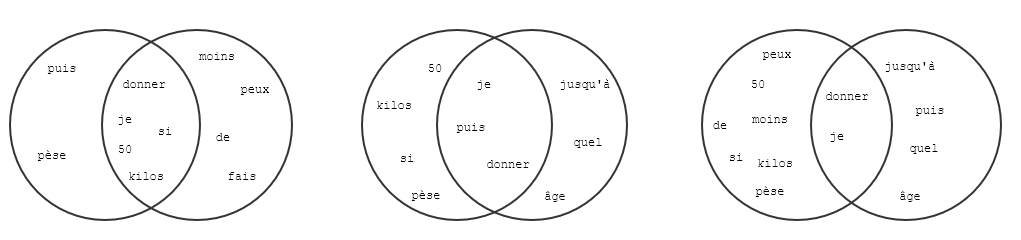
#### Méthodes de similarité textuelle.

### Distance de Jaccard

En 1908, Paul Jaccard s’est intéressé à la distribution de la flore dans les prairies subalpines jurassiennes (Jaccard 1908). Ses résultats ont permis la création de deux nouvelles métriques : l’indice et la distance de Jaccard. Considérons deux ensembles A et B ; l’indice de Jaccard représente le rapport de cardinalité entre l’intersection et l’union de ces ensembles : . Considérons trois ensembles : X = {puis, je, donner, si, je, pèse, 50, kilos}, Y = {je, peux, donner, si, je, fais, moins, de, 50, kilos} et Z = {jusqu’à, quel, âge, puis, je, donner}. On aura alors :

.

On observe ainsi une plus grande similarité entre les ensembles X et Y qu’entre les ensembles X et Z et Y et Z.

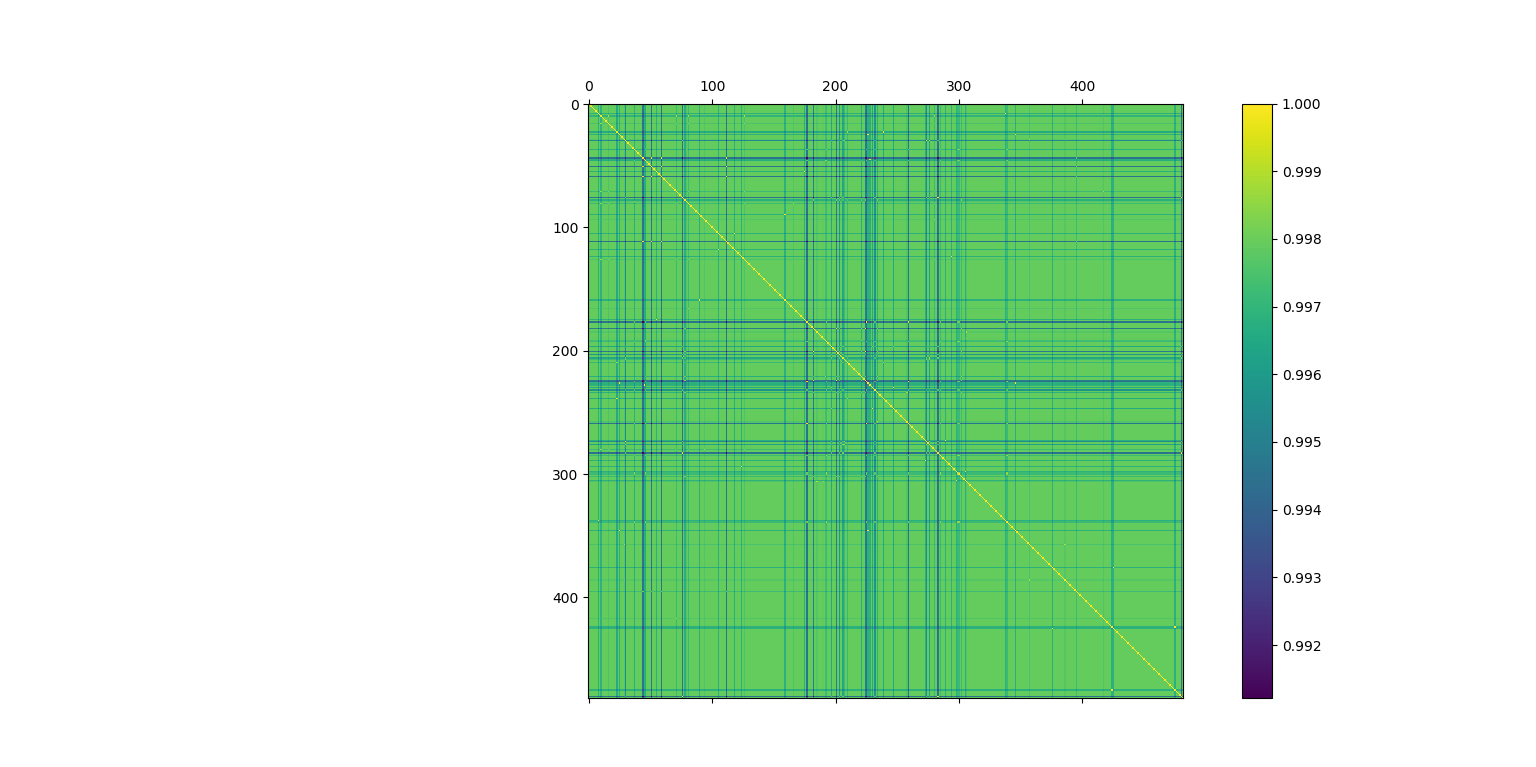


*Figure 4 : diagrammes de Venn de comparaison des ensembles X, Y et Z*

Il est à noter que nous nous sommes contentés de prendre les unités lexicales ‘‘brutes’’, c’est—à-dire sans procéder à aucune lemmatisation ou racinisation ; nous n’avons pas non plus retiré les mots vides. Si tel avait été le cas, les distances entre les ensembles auraient été légèrement réduites, mais n’aurait pas changé les résultats globaux. L’inconvénient principal de la distance de Jaccard est que l’on s’intéresse uniquement à la forme des mots, et non pas à leur sens ; ainsi, des éléments ayant une relation synonymique (comme « anémie ferriprive » et « carence en fer ») augmenteraient la distance entre deux phrases alors qu’ils sont sémantiquement similaires.

### Similarité cosinus

En partant de la méthode tf-idf développée plus haut, on arrive à la vectorisation d’un segment textuel. Pour comparer deux éléments textuels, on peut ainsi mesurer l’angle qui sépare les deux vecteurs ; c’est ce qu’on appelle la similarité cosinus. Dans un espace vectoriel, on calcule généralement la similarité cosinus grâce à une normalisation euclidienne du produit scalaire des deux vecteurs. Le produit scalaire se calcule grâce à, c’est-à-dire la somme de tous les produits des coordonnées et la norme grâce à , c’est-à-dire le produit du vecteur par lui-même*.* Ainsi, la similarité cosinus se calcule de la manière suivante : . Si l’on applique le calcul de similarité cosinus à notre vocabulaire, on obtient le graphique suivant :



On obtient une matrice carrée dont la diagonale principale contient uniquement des 1 (ce qui est évident puisque le mot avec lequel un mot a le plus de similarité est lui-même). Le résultat que nous obtenons n’est cependant pas réellement satisfaisant car les valeurs varient entre 0.99 et 1

#### Mesures de similarité sémantique

#### Analyse des questions

Afin que la machine puisse comprendre ce qui lui est demandé (c’est-à-dire interroger le corpus pour répondre à la question posée par l’utilisateur), il va falloir modéliser la question, c’est-à-dire la faire passer de la langue naturelle à un métalangage compréhensible par le système. En partant de notre hypothèse de filtrage par motif, nous allons tenter en prenant l’exemple spécifique de la question de poids dans le cadre du don du sang, de déterminer les différents motifs possibles.

Q1utilisateur « Je pèse X kilos. Est-ce que je peux donner mon sang ? »

Q2utilisateur « Est-ce que je peux donner mon sang si je pèse X kg ? »

Q3utilisateur « Je peux donner mon sang si je pèse X kg ? »

Q4utilisateur « Si je pèse X kg, je peux donner mon sang ? »

Q5utilisateur « Y a-t-il un poids minimal/maximal pour donner son sang ? »

Q6utilisateur « Est-ce qu’il y a un poids minimal/maximal pour donner son sang ? »

Q7utilisateur « Combien faut-il peser pour donner son sang ? »

Q8utilisateur « Quel est le poids minimal/maximal pour faire le don ? »

V. RESULTATS

….

VI. CONCLUSION

# Annexes et bibliographie

Annexes

**Annexe 1. Taxonomie de** (Li & Roth, 2004)

|  |  |  |  |
| --- | --- | --- | --- |
| Classe | Définition | Classe | Définition |
| **ABBREVIATION** | abréviation | **description** | description de qqch |
| **abb** | abréviation | **manière** | manière dont une action est effectuée |
| **exp** | expression abrégée | **raison** | raisons |
| **ENTITE** | entités | **HUMAIN** | êtres humains |
| **animal** | animaux | **groupe** | groupe ou organisation de personnes |
| **corps** | organes du corps humain | **per** | une personne |
| **couleur** | couleurs | **titre** | le titre d'une personne |
| **création** | inventions, livres, œuvres d'art etc. | **description** | description d'une personne |
| **devise** | monnaies | **ENDROIT** | endroits |
| **mal.med.** | maladies et médicaments | **ville** | villes |
| **évènement** | évènement | **pays** | pays |
| **nourriture** | nourriture | **montagne** | montagnes |
| **instrument** | instrument de musique | **autre** | autres endroits |
| **lang** | langues | **région** | régions |
| **lettre** | lettre de l'alphabet | **NUMERIQUE** | valeurs numériques |
| **autre** | autres entités | **code** | code postal |
| **plante** | plantes | **compte** | quantité de qqch |
| **produit** | produits | **date** | dates |
| **religion** | religions | **distance** | mesures linéaires |
| **sport** | sports | **argent** | prix |
| **substance** | éléments et substances | **ordre** | classement |
| **symbole** | signes et symboles | **autre** | autre nombre |
| **technique** | techniques et méthodes | **période** | durée de qqch |
| **terme** | termes équivalents | **pourcentage** | pourcentage ou fraction |
| **véhicule** | véhicules | **vitesse** | vitesse |
| **mot** | mots avec une spécificité | **temp** | température |
| **DESCRIPTION** | description et concepts abstraits | **taille** | taille, aire et volume |
| **définition** | définition de qqch | **poids** | poids |

**Annexe 2. Liste des sites internet ayant permis la construction du corpus**

[http://www.dondusanglpo.fr/pages/don-du-sang/faq-don-du-sang.html#page3](http://www.dondusanglpo.fr/pages/don-du-sang/faq-don-du-sang.html" \l "page3)

<https://transfusion.be/fr/faq/>

<http://dsb-domerat.chez-alice.fr/faq.html>

<http://sang-pretention.over-blog.com/article-11768161.html>

<http://www.adsb-entzheim.asso.fr/les-dons/faq-sur-le-don-de-sang.html>

<https://www.blood.co.uk/who-can-give-blood/can-i-give-blood/>

<https://www.donateblood.com.au/faq-list>

<https://www.vitalant.org/Resources/FAQs.aspx>

<https://dondesang.efs.sante.fr/faq-sur-le-don>

Bibliographie

Abacha, Asma Ben, and Demner-Fushman Dina. 2016. “Recognizing Question Entailment for Medical Question Answering.” *AMIA ... Annual Symposium Proceedings. AMIA Symposium* 2016 (May): 310–18. http://www.ncbi.nlm.nih.gov/pubmed/28269825%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5333286.

Alliot, J, T Schiex, P Brisset, and F Garcia. 2002. “Intelligence Artificielle et Informatique Théorique (2 Ed.).” *Recherche*.

Antonio, G. A., A. Cisternino, F. Formica, M. Simi, and R. Tommasi. 2001. “PiQASso: Pisa Question Answering System.” *Proceedings of the Text REtrieval Conference*, no. January: 599–607.

Bloomfield, Leonard. 1933. *Language*. New York.

Burke, Robin D, Kristian J Hammond, Vladimir Kulyukin, Steven L Lytinen, Noriko Tomuro, and S Scott Schoenberg. 1997. “Frequently-Asked Question Files: Experiences with the FAQ Finder System.” *AI Magazine* 18 (2): 57–66. http://josquin.cti.depaul.edu/~rburke/pubs/burke-etal-aimag97.ps.

Chomsky, Noah. 2002. *Syntactic Structures*. Edited by De Gruyter Mouton. 2nd ed.

Clark, Alexander, Chris Fox, and Shalom Lappin. 2010. *The Handbook of Computational Linguistics and Natural Language Processing*. Edited by Wiley-Blackwell. http://www.ncbi.nlm.nih.gov/pubmed/15003161%5Cnhttp://cid.oxfordjournals.org/lookup/doi/10.1093/cid/cir991%5Cnhttp://www.scielo.cl/pdf/udecada/v15n26/art06.pdf%5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-84861150233&partnerID=tZOtx3y1.

Cruchet, Sarah, Arnaud Gaudinat, and Célia Boyer. 2008. “Supervised Approach to Recognize Question Type in a QA System for Health.” *Studies in Health Technology and Informatics* 136 (February): 407–12. https://doi.org/10.3233/978-1-58603-864-9-407.

Descles, Jean-Pierre, and Ismail Biskri. 1995. “Logique Combinatoire et Linguistique : Grammaire Catégorielle Combinatoire Applicative.” *Mathématiques et Sciences Humaine* 132: 39–68. https://doi.org/10.1016/B978-0-12-159101-4.50017-6.

Frege, Gottlob. 1892. “On Sense and Reference.” *Zeitschrift Für Philosophie Und Philosophische Kritik*. https://doi.org/10.1093/mind/XCIV.376.526.

———. 1953. *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*. Edited by 2nd. Harper Torchbooks.

Green, Bert, Alice Wolf, Carol Chomsky, and Kenneth Laughery. 1961. “Baseball: An Automatic Question Answerer.” *Proceedings of Western Computing Conference*, 219–24. https://doi.org/10.1017/CBO9781107415324.004.

Halliday, Michael Alexander Kirkwood. 2004. “Introduction to Functional Grammar,” no. December. https://www.functionalmedicine.org/files/library/Intro\_Functional\_Medicine.pdf.

Harris, Zellig. 1968. “Mathematical Structures in Language.”

Jaccard, Paul. 1908. “Nouvelles Recherches Sur La Distribution Florale.” *Bulletin de La Société Vaudoise Des Sciences Naturelles*.

Jeon, Jiwoon, W. Bruce Croft, and Joon Ho Lee. 2006. “Finding Similar Questions in Large Question and Answer Archives,” 84. https://doi.org/10.1145/1099554.1099572.

Jurafsky, Daniel, and James H Martin. 2017. “Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.” https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Katz, Jerrold J., and Jerry A. Fodor. 1963. “The Structure of a Semantic Theory.” *Language* 39 (2): 170. https://doi.org/10.2307/411200.

Langlet, Caroline, and Chloé Clavel. 2015. “Modélisation Des Questions de l’agent Pour l’analyse Des Affects , Jugements et Appréciations de l’utilisateur Dans Les Interactions Humain-Agent,” no. July.

Laurent, Dominique, and Patrick Séguéla. 2004. “QRISTAL, Système de Questions-Réponses.” *TALN 2005*.

Li, Xin, and Dan Roth. 2006. “Learning Question Classifiers: The Role of Semantic Information.” *Natural Language Engineering* 12 (3): 229–49. https://doi.org/10.1017/S1351324905003955.

Ligozat, Anne-laure. 2004. “Système de Question Réponse : Apport de l’analyse Syntaxique Lors de l’extraction de La Réponse.” *RECITAL 2004*.

Luhn, Hans Peter. 1958. “The Automatic Creation of Literature Abstracts.” *IBM Journal of Research and Development* 2 (2): 159–65.

Mel’čuk, Igor A. 1992. “Paraphrase et Lexique: La Théorie Sens-Texte et Le Dictionnaire Explicatif et Combinatoire.” *Dictionnaire Explicatif et Combinatoire Du Français …*. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Paraphrase+et+lexique+:+la+théorie+sens-texte+et+le+dictionnaire+explicatif+et+combinatoire#0.

Mendes, Sara, and Véronique Moriceau. 2004. “L’analyse Des Questions : Intérêts Pour La Génération Des Réponses.” *TALN Workshop Question-Réponse, Fès, 22 Avril 2004*.

Moriceau, Véronique, and Xavier Tannier. 2009. “Apport de La Syntaxe Dans Un Système de Question-Réponse : Étude Du Système FIDJI.” *Architecture*, no. January 2009: 24–26.

Niu, Yun, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. “Answering Clinical Questions with Role Identification.” *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, no. July: 73–80. https://doi.org/10.3115/1118958.1118968.

Omrane, Nouha, Adeline Nazarenko, and Sylvie Szulman. 2009. “Les Entités Nommées : Éléments Pour La Conceptualisation,” no. May: 1–6.

Rinaldi, Fabio, James Dowdall, Gerold Schneider, and Andreas Persidis. 2004. “Answering Questions in the Genomics Domain.” *ACL 2004 Workshop on Question Answering in Restricted Domains*, no. July: 46–53. https://doi.org/10.5167/uzh-19122.

Shannon, Claude E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27 (July 1928): 379–423. https://doi.org/10.1145/584091.584093.

Spärck Jones, Karen. 1972. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval.” *Journal of Documentation* 38: 11–21.

Tesnière, Lucien. 1959. “Eléments de Syntaxe Structurale.”

Wang, Zhiguo, and Abraham Ittycheriah. 2015. “FAQ-Based Question Answering via Word Alignment,” no. 1. http://arxiv.org/abs/1507.02628.

1. https://www.hon.ch/ [↑](#footnote-ref-2)
2. https://www.nltk.org/ [↑](#footnote-ref-3)
3. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [↑](#footnote-ref-4)
4. https://dondesang.efs.sante.fr/ [↑](#footnote-ref-5)
5. Cette traduction n’est pas tout à fait correcte, mais permet de conserver, même en français, l’idée de *coupure.* [↑](#footnote-ref-6)