Le deep learning : un outil pour la didactique du FLE ?

Simona Ruggia

Le *deep learning* : un outil pour la didactique du FLE ?

Simona Ruggia
Maître de conférences en Didactique du FLE
Université Nice Sophia Antipolis (UNS) – Membre Université Côte d'Azur (UCA)
« Bases, Corpus, Langage » UMR 7320 UCA/CNRS/FRANCE

Abstract

Cette contribution démontre l'utilité de l'intelligence artificielle pour la didactique du français langue étrangère, et plus précisément pour l'identification du niveau de textes en français selon le Cadre Européen Commun de Référence pour les Langues (CECRL). Pour ce faire, nous illustrerons les premiers résultats d'une recherche en cours effectuée à l'aide du Text Deconvolution Saliency (TDS) qui implémente l'analyse prédictive du deep learning à l'analyse descriptive grâce à une extraction statistique des saillances qui marquent un changement de niveau selon le CERCL. La comparaison des saillances détectées avec les inventaires des Référentiels pour le français permettra d'une part, d'expliquer l'analyse du TDS et d'autre part, de décrire les caractéristiques de textes en français en fonction de leur niveau.

Mots-clés : *deep learning*, didactique du FLE, niveaux de langue, CERCL.

Abstract

This contribution proves the usefulness of artificial intelligence for didactics of French as a foreign language, and more specifically for the identification of level of French texts according to the Common European Framework of Reference for languages (CEFR). To this end, we will illustrate the first results of a current research carried out using Text Deconvolution Saliency (TDS). The TDS implements the predictive analysis of deep learning to the descriptive analysis thanks to an extraction of the saliencies, which mark a level's change according to the CEFR. The comparison of the detected saliencies with the inventories of the French Référentiels enable on the one end to explain the TDS's analysis, and on the other end to describe the characteristics of French texts depending of their level.

Keywords : deep learning, didactics of French as a foreign language, language levels, CEFR.

Introduction

Le *Cadre Européen Commun de Référence pour les langues* (CERCL)[1] a permis, entre autres, d'établir les niveaux de compétences, allant de A1 à C2, qu'un apprenant peut acquérir lors de son apprentissage d'une langue étrangère. Depuis sa publication de nombreux outils didactiques et pédagogiques ont vu le jour afin de faciliter la tâche des apprenants et des enseignants.
C'est dans cette lignée que notre étude propose une méthode novatrice permettant d'identifier le niveau d'un texte, en d'autres mots de « toute séquence discursive (orale et/ou écrite) »[2] grâce au *deep learning*.
Dans le cadre de notre recherche, nous avons fait appel au *Text Deconvolution Saliency* (TDS)[3] afin d'extraire et décrire les caractéristiques linguistiques qui marquent un changement de niveau du CECRL pour les textes en français. Pour ce faire, notre recherche s'est appuyée sur un corpus échantillonné et constitué de textes oraux extraits de plusieurs ensembles pédagogiques de français langue étrangère (FLE) de niveau A1 et B2[4].
Cette contribution illustrera les premiers résultats d'une recherche en cours sur l'efficacité du *deep learning*, et plus particulièrement du TDS lors de la reconnaissance du niveau d'un texte en français. Elle vise également à analyser et expliquer les résultats du TDS en les comparant aux inventaires des Référentiels pour le français.

Le *deep learning* au service de la didactique du FLE

La question des niveaux de langue, de leur caractérisation et de leur maîtrise est une question centrale de la didactique du FLE et surplombe donc l'analyse automatique des corpus.
Bien que ces niveaux soient décrits de manière détaillée par le CERCL, force est de constater qu'évaluer le niveau d'un texte authentique ou fabriqué destiné à l'apprentissage du FLE ou encore d'une production orale ou écrite d'un apprenant nécessite un degré d'expertise très élevé de la part d'un enseignant. C'est pour cette raison que nous avons décidé d'exploiter les fonctionnalités du *deep learning* au service de la didactique du FLE. Notre projet de recherche s'est construit autour de l'hypothèse suivante : le *Text Deconvolution Saliency* est capable d'extraire les caractéristiques de textes en français et plus précisément il est capable d'extraire les saillances qui marquent un changement de niveau selon le CERCL.

Le *deep learning*

Le *deep learning* est un type d'intelligence artificielle, qui est une branche de l'informatique fondamentale permettant de simuler des comportements du cerveau humain. Plus précisément, le *deep learning* est une technologie d'apprentissage et de classification basée sur des réseaux de neurones artificiels permettant d'apprendre à reconnaître des images, des voix, des textes.
Les premières études sur les réseaux de neurones remontent à 1943 (McCulloch : neurophysiologiste et Pitts : logicien), néanmoins le *deep learning* a été formalisé en 2007 à partir de nouvelles architectures de réseaux de neurones. Mais c'est seulement, depuis 2012 que les recherches sur le *deep learning* ont abouti à l'élaboration de méthodes efficaces permettant d'avoir une plus grande précision au niveau de la reconnaissance grâce à l'élaboration de nouveaux algorithmes, à la réalisation de grands corpus appelés aussi *big data*, mais surtout grâce à la puissance des machines.
Le *deep learning* est souvent comparé à une boîte noire car le système fait une analyse et une reconnaissance mais il n'explicite pas les critères utilisés, en d'autres termes on ne sait pas comment le système prend sa décision. C'est donc afin d'expliciter ces critères que le TDS a été développé par L. Vanni, ingénieur du CNRS, au sein du laboratoire Bases, Corpus et Langage[5]. Dans cette perspective,

[1] Conseil de l'Europe (2001), *Cadre Européen Commun de Référence pour les Langues*, Didier, Paris.
[2] Ivi, p.15.
[3] L. Vanni *et al.* (2018), *Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis*, in 56th Annual Meeting of the Association for Computational Linguistics, jul 2018, Melbourne [hal-01804310].
[4] Voir bibliographie : Ouvrages de FLE utilisées pour le corpus.
[5] UMR 7320 UCA/CNRS/FRANCE.

le TDS implémente l'analyse prédictive du deep learning à l'analyse descriptive grâce à une extraction statistique des passages-clefs « avec une évaluation de leur pertinence interprétative »[6].
La recherche sur l'analyse de données textuelles est au cœur de ce laboratoire depuis plusieurs années et a donné lieu à de nombreuses publications ainsi qu'à la création du logiciel Hyperbase par E. Brunet, dont la première version remonte à 1989 et qui a été ensuite développé par L. Vanni pour la version Hyperbase web que nous avons également utilisé pour notre recherche.

Le *Text Deconvolution Saliency* et l'Analyse des Données Textuelles (ADT)

Le TDS est un système d'apprentissage profond conçu grâce à l'élaboration d'un algorithme : le *Query-By-Dropout-Committee* (QBDC) qui « sélectionne itérativement les échantillons les plus pertinents pour être ajoutés à l'ensemble d'entraînement afin que le modèle soit amélioré de façon optimale »[7]. Le TDS est capable d'extraire les caractéristiques qui donnent une empreinte unique du texte, en d'autres termes c'est une application qui détecte les saillances du texte à différents niveaux linguistiques (lexique, grammaire, morphosyntaxe …) grâce au travail préalable de lemmatisation. Mais qu'entend-on par saillances ? Ces dernières correspondent aux occurrences que le système juge être les plus importantes et qui permettent la reconnaissance.
La recherche sur le *deep learning* se nourrit de celle sur l'Analyse des données textuelles et vice-versa. Dans ce sens, la notion de passage tel qu'il a été défini par F. Rastier, à savoir « un extrait, entre deux blancs s'il s'agit d'une chaine de caractères ; entre deux pauses ou ponctuations, s'il s'agit d'une période »[8] représente un concept heuristique de notre analyse. La notion de passage part de la « grandeur » du texte pour en définir sa « grandeur locale » qui peut correspondre à « un signe, à une phrase, ou par exemple à un paragraphe »[9]. De plus, le passage est « un morceau de texte jugé suffisamment parlant […] pour prétendre rendre compte d'un texte »[10]. C'est dans ce sens que nous assignons au TDS la tâche de repérer les passages-clefs en considérant un passage-clef comme : « une unité de surcroit textométrique ; c'est-à-dire une unité dont la pertinence est calculable et l'extraction automatique »[11].
La recherche sur le TDS est certes très récente mais elle a déjà donné des résultats satisfaisants, notamment avec des corpus en français de discours politiques et de textes littéraires mais aussi avec des corpus en anglais et en latin.
Dans le cadre d'un projet IDEX[JEDI] Académie 5 de l'Université Côte d'Azur deux plateformes ont été créées : la première « Mesure du discours »[12] : est un observatoire du discours politique français outillé par la statistique. La deuxième : « Deeptext »[13] permet aux utilisateurs de tester la fonction de prédiction de reconnaissance du lexique des sentiments d'un texte en anglais, de l'auteur d'un texte en latin et de la tendance politique (gauche ou droite) d'un texte en français.

[6] L. Vanni, D. Mayaffre, D. Longrée (2018) *ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables*, JADT 2018, in Actes des 14èmes Journées internationales d'Analyse statistique des Données textuelles, Rome, Italie p.460.
[7] M. Ducoffe *et al.* (2016), *Machine learning under the light of phraseology expertise : use case of presidential speeches, de Gaulle – Hollande (1958-2016)*, in JADT 2016, Actes des 13èmes Journées internationales d'Analyse statistique des Données textuelles, vol.1, Nice, p.158.
[8] F. Rastier (2007), *Passages*, in Corpus, n°6, p.30.
[9] *Ibidem*,
[10] L. Vanni, D. Mayaffre, D. Longrée (2018), *ADT et deep learning, regards croisées* …, cit., p.461.
[11] Ivi, p.461.
[12] http://mesure-du-discours.unice.fr/ : Partenaires scientifiques : UMR 7320 : Bases, Corpus, Langage (équipe logométrie), I3S (équipe SPARKS). Comité scientifique : Damon Mayaffre, Laurent Vanni, Magali Guaresi, Camille Bouzereau, Fréderic Précioso, Mélanie Ducoffe, Dominique Longrée, Sylvie Mellet.
[13] http://deeptext.unice.fr. Projet chapoté par le même comité scientifique que *supra*.

Méthodologie et corpus

Notre recherche, dont nous présentons ici les premiers résultats, vise à créer un outil d'analyse, de classification et d'identification du niveau de textes en français selon les niveaux du CERCL. Pour ce faire, nous avons tout d'abord établi un corpus d'apprentissage qui va être entraîné par le TDS et qui permettra par la suite la reconnaissance des niveaux. Afin d'expliciter les résultats du TDS et de pouvoir vérifier notre hypothèse de départ, nous avons comparé les saillances détectées par le TDS avec les inventaires des Référentiels pour le français pour les niveaux A1 et B2[14]. Le but de ces ouvrages prépédagogiques est de « transposer les descriptions du *Cadre*, établies en termes de compétences (ou éléments de compétences) et de niveaux (ou degrés) de maîtrise dans une compétence en inventaires de signes linguistiques »[15].
Le corpus actuel comprend des textes oraux en français de niveau A1 et B2 extraits de divers manuels de FLE, mais notre corpus définitif couvrira tous les niveaux de A1 à C2 et sera constitué de séquences discursives orales et écrites. Il est important de préciser que pour qu'un corpus d'apprentissage soit efficace, il doit comporter au moins 100000 occurrences :

NIVEAU CECRL	OCCURENCES	MOTS	PONCTUATION
A1	101923	83540	18383
B2	97961	84390	13571

Le choix du corpus d'apprentissage a été effectué selon des critères bien définis, les textes devaient être issus de :
- manuels de méthodes ou ensembles pédagogiques (autrement dit des ouvrages conçus pour l'apprentissage de la langue française et plus précisément qui permettent d'acquérir une compétence à communiquer langagièrement en français) qui s'inscrivent dans l'approche actionnelle du CECRL,
- manuels de méthodes couvrant plusieurs niveaux,
- manuels de méthodes récentes,
- ouvrages conçus pour la préparation aux diplômes du DELF et DALF[16].

Le travail de préparation du corpus a été très long et minutieux. La plupart des textes ont été scannés, une partie copiée à partir des guides pédagogiques en version pdf pour ensuite créer des fichiers au format txt[17]. Il a fallu effectuer un toilettage afin d'éliminer les éléments (fautes, symboles …) qui pourraient entraver la reconnaissance. Nous avons également modifié et uniformisé les signes de ponctuation, ainsi au sein des tours de parole les points ont été remplacés par des points virgules, pour que les points se trouvent seulement à la fin des tours de paroles.
Les tableaux *infra* illustrent les détails de chaque sous-partie du corpus. On remarquera que pour le niveau A1 onze ouvrages ont été utilisés alors que pour le niveau B2 seulement cinq car les textes de ce niveau sont beaucoup plus longs. En ce qui concerne le nom du fichier, il comporte quatre mots séparés par un underscore, car ces mots correspondent aux métadonnées du corpus, de cette manière à l'aide de la plateforme Hyperbase web[18], on peut varier les critères d'analyse du corpus, critères qui correspondent donc aux métadonnées. Nous tenons aussi à préciser que nous avons distingué en deux catégories les séquences discursives à l'oral : à savoir les interactions et les monologues conformément au classement du CERCL.

[14] J.C. Beacco, S. Bouquet, R. Porquier (2007), *Niveau A1 pour le français, un référentiel*, Didier, Paris.
J.C. Beacco, R. Porquier (2004), *Niveau B2 pour le français, un référentiel*, Didier, Paris.
[15] J.C. Beacco, R. Porquier (2004), *Niveau A1 pour le français…*, cit., p.9.
[16] Diplôme d'Etude en Langue Française, Diplôme Approfondi de Langue Française.
[17] Il s'agit du format nécessaire pour créer la base de données.
[18] « Hyperbase combine deux types de fonctions, documentaires et statistiques, qui permettent à l'analyste de décrire, caractériser, classer et interpréter les textes », <http://hyperbase.unice.fr>.

			Nom du fichier	occurrences	mots	ponctuation
A1	ORAL	2018	A1_ORAL_EDITO_INTERACTION	9073	7525	1548
			A1_ORAL_EDITO_MONOLOGUE	3534	3004	530
		2018	A1_ORAL_DEFI_INTERACTION	2916	2435	481
			A1_ORAL_DEFI_MONOLOGUE	2158	1853	305
		2017	A1_ORAL_COSMOPOLITE_INTERACTION	8506	6992	1514
			A1_ORAL_COSMOPOLITE_MONOLOGUE	6334	5451	883
		2016	A1_ORAL_TENDANCES_INTERACTION	6878	5596	1282
			A1_ORAL_TENDANCES_MONOLOGUE	1921	1646	275
		2014	A1_ORAL_TOTEM_INTERACTION	2338	1886	452
			A1_ORAL_TOTEM_MONOLOGUE	738	609	129
		2012	A1_ORAL_MOBILE_INTERACTION	8093	6507	1586
			A1_ORAL_MOBILE_MONOLOGUE	1608	1285	323
		2013	A1_ORAL_ECHO_INTERACTION	3881	3107	774
			A1_ORAL_ECHO_MONOLOGUE	1763	1455	308
		2009	A1_ORAL_LENOUVEAUTAXI_INTERACTION	6980	5503	1477
			A1_ORAL_LENOUVEAUTAXI_MONOLOGUE	1914	1549	365
		2007	A1_ORAL_ICI_INTERACTION	4375	3628	747
			A1_ORAL_ICI_MONOLOGUE	1782	1535	247
		2007	A1_ORAL_ALORS_INTERACTION	3518	2738	780
			A1_ORAL_ALORS_MONOLOGUE	660	562	98
		2006	A1_ORAL_ALTER EGO_INTERACTION	7276	5875	1401
			A1_ORAL_ALTER EGO_MONOLOGUE	1273	1050	223
		2005	A1_ORAL_DELF150activites_INTERACTION	5843	4611	1232
			A1_ORAL_DELF150activites_MONOLOGUE	3304	2756	548
		2005	A1_ORAL_CECR_INTERACTION	2590	2134	456
			A1_ORAL_CECR_MONOLOGUE	2667	2248	419
			Total	**101923**	**83540**	**18383**

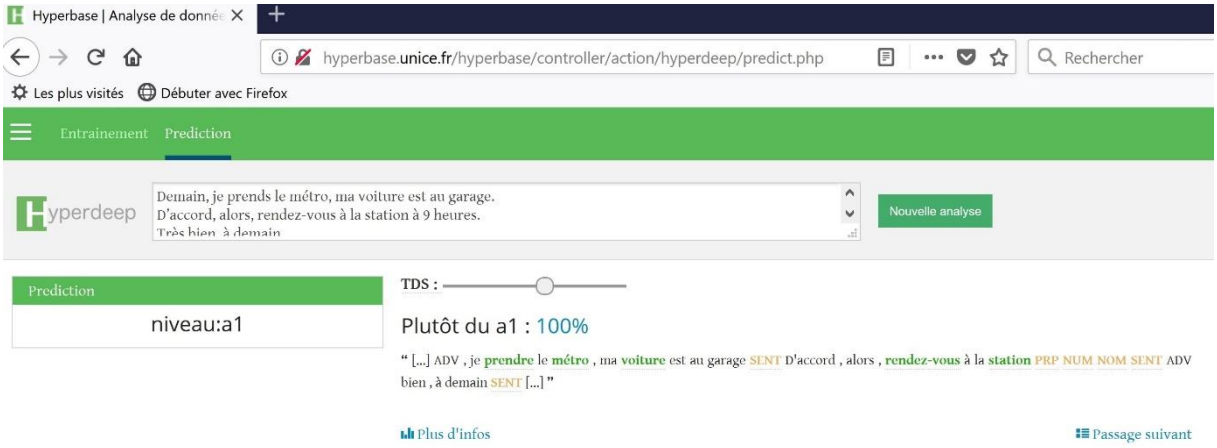
			Nom du fichier	occurrences	mots	ponctuation
B2	ORAL	2017	B2_ORAL_ECHO_INTERACTION	13751	11959	1792
			B2_ORAL_ECHO_MONOLOGUE	4984	4391	593
		2017	B2_ORAL_TENDANCES_INTERACTION	15893	13757	2136
			B2_ORAL_TENDANCES_MONOLOGUE	5283	4573	710
		2016	B2_ORAL_EDITO_INTERACTION	13555	11424	2131
			B2_ORAL_EDITO_MONOLOGUE	4716	4165	551
		2007	B2_ORAL_CECR_INTERACTION	13102	11141	1961
			B2_ORAL_CECR_MONOLOGUE	4450	3876	574
		2007	B2_ORAL_ALTER EGO_INTERACTION	18891	16175	2716
			B2_ORAL_ALTER EGO_MONOLOGUE	3336	2929	407
			Total	**97961**	**84390**	**13571**

Expérience

Après avoir préparé et entraîné[19] le corpus d'apprentissage décrit *supra*, nous avons pu tester par le biais de la plateforme Hyperbase[20], la reconnaissance du niveau d'un certain nombre de textes. A cette fin, nous avons préparé un deuxième corpus contenant plusieurs textes de niveau A1 et B2 attestés, autrement dit extraits d'autres manuels de FLE qui n'ont pas été intégrés au corpus d'apprentissage.
Comme nous l'avons déjà précisé, la plateforme Hyperbase web offre plusieurs fonctionnalités permettant l'analyse des données textuelles ainsi que la fonction « hyperdeep prédiction » que nous avons utilisée pour notre expérience. Afin de présenter en détail les résultats de la prédiction, nous ne reproduirons ici qu'un exemple, à l'aide du dialogue suivant[21] :

> Demain, je prends le métro, ma voiture est au garage.
> D'accord, alors, rendez-vous à la station à 9 heures.
> Très bien, à demain.

L'image *infra* permet de visualiser les résultats de l'analyse : le niveau de cette interaction est bien A1 avec un score de 100%.



Comment le TDS a-t-il reconnu le niveau ? Il faut savoir que tous les mots ont un indice et ont du sens pour la décision finale. La visualisation du seuil de reconnaissance se fait en déplaçant le curseur[22], quand ce dernier est plus à droite on visualise les saillances les plus fortes du texte, en d'autres termes les occurrences les plus importantes pour la reconnaissance.
L'analyse effectuée est liée à la lemmatisation préalable du texte[23] et les résultats sont visibles grâce aux couleurs attribuées à certaines occurrences. Ainsi, le bleu indique qu'il s'agit d'une occurrence que le logiciel reconnaît en tant que mot, il reconnait donc la forme graphique, l'orange indique la catégorie grammaticale et des sigles indiquent le type de catégorie (nom propre, verbe …), enfin le vert indique qu'il s'agit d'un lemme[24].

---

[19] Nous tenons à remercier L. Vanni qui s'est chargé de l'entrainement du corpus et sans qui cette expérience n'aurait pas pu voir le jour.
[20] Sur la page d'accueil de la plateforme Hyperbase web l'utilisateur peut créer une nouvelle base de données, accéder à une base existante mais privée et donc nécessitant un mot de passe, ou accéder à la liste des bases de données publiques.
[21] D. Clément-Rodriguez (2018), *ABC DELF A1 200 exercices*, Paris, CLE International.
[22] Le curseur se trouve au-dessus du résultat de la prédiction, ici « plutôt du A1 : 100% ».
[23] La lemmatisation a été réalisé à l'aide de TREETAGGER.
[24] Dans l'image *supra* les occurrences « prendre », « métro », « voiture », « rendez-vous », « station » apparaissent en vert et en orange les sigles : SENT = fin de phrase, PRP = déterminant, NUM = numéral, NOM = nom, le sigle ADV ici en noir = adverbe.

Analyse des résultats de la prédiction du *deep learning* et de la description par le TDS

Afin d'expliquer les résultats de la prédiction du TDS et de démontrer l'intérêt de celui-ci pour la recherche en didactique du FLE, nous avons dans un premier temps mis en correspondance les saillances détectées par le TDS avec les inventaires des Référentiels pour le français[25].
Comme nous l'avons précisé les saillances sont observables grâce aux couleurs et sigles attribués par le TDS :

mais elles peuvent aussi être visualisées grâce au graphique *infra* qui hiérarchise les saillances du texte en montrant le taux d'activation de ces dernières pour la prédiction. En ce qui concerne le dialogue en question, c'est le chiffre 9 qui a le taux d'activation le plus élevé.

Commençons donc notre analyse par le chiffre « 9 ». Le chapitre 4 « Notions générales » du Référentiel A1 qui « part des notions générales, c'est-à-dire de catégories sémantiques relatives à l'existence, l'espace, le temps, la quantité … pour répertorier et classer les unités lexicales correspondantes »[26] mentionne les nombres et les numéros :

> **4.2. QUANTITES**
> **4.2.1. Nombres**
> **Noms** nombre, numéro[27]

Dans ce chapitre la notion de temps est aussi abordée et, parmi les noms repérés par le TDS, nous trouvons sur le graphique l'occurrence « heure » de notre dialogue :

> **4.4. TEMPS**
> **4.4.1. Divisions du temps**
> **Noms** temps, moment, année, printemps, été, automne, hiver, mois, janvier, février, mars, avril, mai, juin, juillet, août, septembre, octobre, novembre, décembre, jour, semaine, week-end, lundi, mardi, mercredi, jeudi, vendredi, samedi, dimanche, jour, nuit, matin, après-midi, soir, seconde, minute, heure, midi, minuit[28]

Pour ce qui concerne les occurrences « métro, voiture et station » elles apparaissent dans le chapitre 6 « Notions spécifiques » :

> **6.9. TRANSPORT ET VOYAGES**
> **6.9.2. Moyens de transport publics**
> **Noms** train, métro, autobus, tramway, taxi, avion, bateau, place
>
> Départ et arrivée
> **Noms** station, gare, quai, voie, aéroport, porte d'embarquement, enregistrement, retard
>
> **6.9.3. Moyens de transport privés**
> **Noms** auto, voiture, moto, scooter, vélo, bicyclette[29]

[25] Les Référentiels sont organisés en 10 chapitres : 1) Structure du Niveau, 2) Spécifications générales du Niveau - de la compétence de communication au répertoire discursif, 3) Fonctions, 4) Notions générales, 5) Grammaire : morphologie et structures des énoncés et des phrases, 6) Notions spécifiques, 7) Matière sonore, 8) Matière graphique, 9) Compétences culturelles, 10) Stratégies d'apprentissage.
[26] J.C., Beacco, R. Porquier. (2007), *Niveau A1 pour le français …,* cit., p.77.
[27] Ivi. p.80), C'est nous qui soulignons ici et dans les extraits suivants du Référentiel.
[28] Ivi, p.84.
[29] Ivi p.115.

C'est aussi dans ce chapitre qui est mentionné le verbe « prendre » :

> **6.1. L'ETRE HUMAIN**
>
> **6.1.7. Opérations manuelles**
> **Verbes** prendre, mettre, tenir, ouvrir, fermer[30]

Le verbe « prendre » a été mis en évidence par le système et étiqueté en vert en tant que lemme, en ce sens, l'analyse montre que la caractéristique la plus importante qui a permis la reconnaissance de cette occurrence est le verbe à l'infinitif. A ce propos, nous retrouvons dans le chapitre 5 « Grammaire : morphologie et structures des énoncés et des phrases » les temps et modes verbales qui caractérisent un texte produit par un apprenant de A1 :

> **5.1. Morphologie**
> **5.1.1. Morphologie des verbes**
> **5.1.1.1. Flexion : nombres et personnes**
> Au niveau A1, l'apprenant/utilisateur est capable d'identifier les formes flexionnelles (nombres et personnes) […] d'utiliser des verbes au présent de l'indicatif, à l'infinitif, à l'impératif, ainsi que, de manière isolé et sporadique, au passé composé […] et, éventuellement à l'imparfait pour un nombre très limité de verbes (*il/c'était, il (y) avait, il faisait* …)[31]

Il apparaît donc que les saillances détectées par le TDS correspondent aux caractéristiques des textes telles que décrites par le Référentiel du niveau A1.

Pour ce qui est du niveau B2, le taux de score de reconnaissance est aussi très élevé[32] : 97,37%, comme on peut le voir sur l'image *infra* :

[30] Ivi, p.110.
[31] Ivi, p.97.
[32] La prédiction a été effectuée avec un texte beaucoup plus long que celui du niveau A1, à savoir trois exemples de monologues, dont nous ne reproduisons ici qu'une partie « La révolution que les découvertes biologiques de Louis Pasteur ont entraînée a paradoxalement provoqué, dans l'inconscient collectif, une véritable peur, une phobie des bactéries, qui du coup sont devenues les responsables de tous nos maux et presque toutes nos maladies ; s'il est vrai que d'un point de vue strictement sanitaire et médical, certaines de ces bactéries ont été à l'origine des plus grandes maladies mortelles, faut-il cependant craindre que … »
M.-L. Parizet (2018), *ABC DELF B2 200 exercices*, Paris, CLE International.

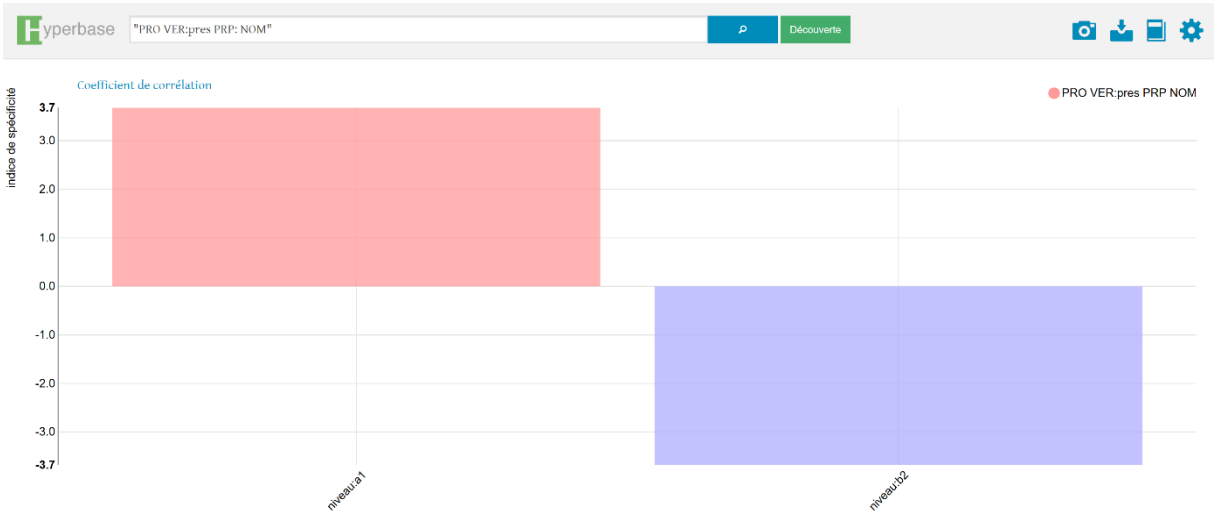
Notre réflexion sur les résultats de la prédiction du TDS et sur les caractéristiques des niveaux des textes, fait également appel à la notion de passage-clef, telle que décrite *supra*, car cette unité textométrique nous semble particulièrement pertinente dans le cadre de notre recherche.
Ainsi, à l'aide de la plateforme Hyperbase web, nous avons cherché la distribution des passages-clefs sélectionnés par le TDS pour la reconnaissance du niveau A1 dans notre corpus d'apprentissage. A titre d'exemple, nous illustrerons la distribution de deux passages-clés détectés dans le texte A1 préalablement analysé :

1) je prends le métro

2) ma voiture est au garage

Le graphique de la distribution du 1er passage-clé : « je prends le métro » indique l'indice de spécificité de cette unité linguistique constituée d'un pronom, un verbe au présent, un déterminant et un nom[33] pour les textes de niveau A1 par rapport aux textes de niveau B2 :

L'indice de spécificité de la distribution du 2e passage-clé : « ma voiture est au garage » est aussi très élevé :



[33] Afin d'effectuer une analyse de passages-clés avec hyperbase web, l'utilisateur doit les « traduire » en utilisant les codes de lemmatisation correspondant aux différentes parties du discours, tels qu'ils ont été définis par cette plateforme. Ainsi le passage-clé : « je prends le métro » correspond à : « PRO VER:pres DET NOM ».

Ces graphiques montrent que les passages-clés sélectionnés caractérisent fortement les textes de niveau A1, contrairement à ceux du niveau B2.
Mais pouvons-nous affirmer que ces passages-clés sont vraiment typiques du niveau A1 ? Afin de corroborer les résultats du TDS et de l'ADT, nous avons cherché ces passages-clés dans le Référentiel pour le niveau A1. Et il s'avère que ces deux unités linguistiques sont recensées parmi les contenus du niveau A1 « posés comme devant être maîtrisés dans les activités de réception/compréhension orale et/ou écrite ».[34] Plus précisément, ces passages-clés sont listés dans le chapitre 5 « Grammaire : morphologie et structures des énoncés et des phrases » :

5.2. Structures de phrase simple
5.2.3. GN GV - : constructions verbales
5.2.3.1. GN V (GN/adj)
GN V Le printemps arrive
GN V GN J'attends le train.
GN Vêtre Adj Tu es content ?[35]

Et dans le chapitre 3 « Fonctions » :

3.1. Interagir à propos d'informations
3.1.8. Répondre à une demande d'information
3.1.8.2. … en donnant des informations
3.1.8.2.2. … sur le lieu
Adv… Ici.
Gprép. A côté de la poste.
P. Elle est là.[36]

Conclusion - Perspectives

Ce travail exploratoire a permis de mettre en évidence les atouts de l'intelligence artificielle pour la didactique du FLE et de valider notre hypothèse de départ.
Les premiers résultats sur l'utilisation du *deep learning* et du *Text Deconvolution Saliency* pour la reconnaissance et la description des niveaux de textes en français selon le CECRL, nous incitent à poursuivre notre recherche qui vise la création d'une plateforme permettant l'identification et la description des caractéristiques de tous les niveaux de textes (de A1 à C2) en français aussi bien à l'oral qu'à l'écrit. Cet outil sera destiné aux divers acteurs du FLE, à savoir : les concepteurs de programmes, les concepteurs de manuels, les concepteurs de diplômes et certifications, les évaluateurs, les enseignants et les apprenants.

Bibliographie

Didactique et logométrie

Beacco J.C., Bouquet S., Porquier R. (2004), *Niveau B2 pour le français, un référentiel*, Didier, Paris.

Beacco J.C., Porquier R. (2007), *Niveau A1 pour le français, un référentiel*, Didier, Paris.

Conseil de l'Europe (2001), *Cadre Européen Commun de Référence pour les langues : apprendre, enseigner, évaluer*, Didier, Paris.

[34] J.C. Beacco, R. Porquier (2007), *Niveau A1 pour le français*…, cit., p.15.
[35] Ivi, p.98.
GN : groupe nominal. V : verbe. Vêtre : verbe être. Adj : adjectif.
[36] Ivi, p.58.
Adv : adverbe. G prép : groupe prépositionnel. P : proposition.

Ducoffe M., Precioso F., Arthut A., Mayaffre D., Lavigne F., Vanni L. (2016), *Machine learning under the light of phraseology expertise : use case of presidential speeches, de Gaulle – Hollande* (1958-2016), JADT 2016, in *Actes des* 13èmes *Journées internationales d'Analyse statistique des Données textuelles*, vol.1, Nice, France, pp.157-168. Consultable à l'adresse : <http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/86038/86038.pdf>.

Rastier F. (2007), *Passages*, in Corpus, n°6, pp.25-54.

Vanni L., Ducoffe M., Mayaffre D., Precioso F., Longrée D., *et al*. (2018), *Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis*, 56th Annual Meeting of the Association for Computational Linguistics, jul 2018, Melbourne, France. [hal-01804310]

Vanni L., Mayaffre D., Longrée D. (2018), *ADT et deep learning, regards croisés. Phrases-clefs, motif et nouveaux observables*, JADT 2018, in Actes des 14èmes Journées internationales d'Analyse statistique des Données textuelles, Rome, Italie, pp.459-466. Consultable en ligne à l'adresse : <http://lexicometrica.univ-paris3.fr/jadt/JADT2018/actes-jadt18.pdf>.

Ouvrages de FLE utilisés pour le corpus

Abry D., Fert C., Parpette C., Stauber J., Soria M., Borg S. (2007), *Ici méthode de français A1*, Paris, CLE International.

Alcaraz M., Braud C., Calvez A., Cornuau G., Jacob A., Pinson C., Vidal S. (2016), *Edito méthode français, niveau A1*, Paris, Didier.

Berthet A., Hugot C., Kizirian V., Sampsonis B., Waendendries M. (2006), *Alter ego méthode de français A1*, Paris, Hachette.

Capelle G., Menand R. (2009), *Le nouveau Taxi méthode de français A1*, Paris, Hachette.

Chahi F., Denyer M., Gloanec A., Briet G., Neuenschwander V., Fouillet R. (2018), *Défi méthode de français A1*, Paris, Maison des langues.

Clément-Rodriguez D. (2018), *ABC DELF A1 200 exercices*, Paris, CLE International.

Di Giura M., Beacco J.-C. (2007), *Alors méthode de français fondée sur l'approche par compétences A1*, Paris, Didier.

Dollez C., Pons S. (2007), *Alter ego méthode de français B2*, Paris, Hachette.

Girardet J., Gibbe C. (2017), *Echo 2e édition méthode de français B2*, Paris, CLE International.

Girardet J., Pécheur J. (2013), *Echo 2e édition méthode de français A1*, Paris, CLE International.

Girardet J., Pécheur J., Gibbe C., Parizet M.-L. (2016), *Tendances méthode de français A1*, Paris, CLE International.

Girardet J., Pécheur J., Gibbe C., Parizet M.-L. (2017), *Tendances méthode de français B2*, Paris, CLE International.

Heu E., Mabilat J.-J. (2016), *Edito méthode de français B2*, Paris, Didier.

Hirschsprung N., Tricot T. (2017), *Cosmopolite méthode de français A1*, Paris, Hachette.

Lescure R., Gadet E., Vey P., Bloomfield A., Daill E. (2005), *DELF A1 150 activités*, Paris, CLE International.

Lopes M.-J., Le Bougnec J.-T. (2014), *Totem méthode de français A1*, Paris, Hachette.

Parizet M.-L. (2018), *ABC DELF B2 200 exercices*, Paris, CLE International.

Parizet M.-L., Grandet E., Corsain M. (2005), *Activités pour le Cadre Européen Commun de Référence*, Paris, CLE International.

Reboul A., Boulinguez A.-C., Fouquet G. (2012), *Mobile méthode de français A1*, Paris, Didier.

Sitographie

Mesure du discours :
Discours présidentiels français de 1958 à aujourd'hui. Observatoire du discours politique français. Méthodes logométriques & deep learning.
http://mesure-du-discours.unice.fr

DEEP TEXT :
Technical demonstration powered by hyperbase.unice.fr linguistics web tool.
http://deeptext.unice.fr

Hyperbase web :
Logiciel d'analyse de données textuelles.
http://hyperbase.unice.fr